НЕ ТОЛЬКО РАЗМЕР ИМЕЕТ ЗНАЧЕНИЕ: АСПЕКТЫ СОЗДАНИЯ ЧАСТОТНЫХ СЛОВАРЕЙ НА ОСНОВЕ КОРПУСОВ

С.А. ШАРОВ
s.sharoff@leeds.ac.uk
канд. физ.-мат. наук, доцент Университета Лидса
Лидс, Великобритания

Ключевые слова:
надежность частотных данных, всплески частот, жанровое разнообразие, иллюстративные примеры, оценка сложности

DOI: 10.37632/PI.2020.283.6.002

В статье рассматриваются проблемы создания частотных словарей для преподавания языка с учетом таких параметров, как источники корпусов, собственно частотность слова, зависимость от длины документов, тематическое и жанровое разнообразие корпусов. Приводятся примеры проблем с частотными списками и даются рекомендации для практического применения частотных словарей. Отмечается, что помимо размера корпуса на содержание частотных словарей влияют слова, популярные внутри длинных документов, поскольку они приводят к выбросам частот, а также соответствие тем и жанров, представленных в корпусе, целям обучения, так как корпуса из разных предметных областей и жанров могут радикально отличаться друг от друга.

Правильно организованное преподавание иностранных языков предполагает построение процесса обучения от более важных слов и конструкций к менее важным. Этапы обучения фиксируются в тестах, определяющих уровень владения языком (например, ТРКИ или CEFR), и связанных с ними лексических и грамматических минимумах. Один из способов определения того, что является важным, заключается в учете информации о частоте лингвистических явлений: наиболее важно ядро языка, которое необходимо для большинства актов коммуникации. Такой подход к ранжированию по важности приводит к активному использованию информации о частоте конструкций разного вида, поскольку предполагается, что такие данные естественным образом ранжируются по степени их принадлежности ядру языка. Например, для удовлетворительного понимания текста студентам требуется знание примерно 80–90% слов этого текста [16]. В свою очередь, для достижения 80% покрытия требуется знание около 5 тыс. часто встречающихся слов [5, 17]. Преподавание, структурированное с учетом частот слов, позволяет избежать волюнтаризма и построить педагогический процесс в соответствии с лингвистической теорией, которая в настоящее время также основана на использовании больших объемов данных в виде корпусов текстов.

Несмотря на разумность этой точки зрения, в использовании больших объемов данных есть много подводных камней. Необходимо понимать, что большие данные сами по себе не панацея для получения качественного частотного словника. В статье обсуждаются такие аспекты, как:

- надежность частотных данных, включая влияние размера корпуса, относительных размеров документов и частотных выбросов;
- отражение вариативности языка в корпусах и частотных списках, включая тематическое и жанровое разнообразие;
- отбор подходящих иллюстративных примеров.

Чтобы понять текущее состояние дел, стоит взглянуть на историю создания частотных словарей. Рост интереса к использованию

научных знаний и статистики в педагогике начиная с 1930-х гг. привел к появлению частотных словарей для преподавания. Видимо, первым таким словарем для русского языка стал словарь Йоссельсона, изданный для преподавания русского языка в США [14]. Затем появился словарь Штейнфельд, созданный в Эстонии с целью определения лексического минимума «нерусских» (как сказано в публикации) в средних школах [7]. Наконец, на полноценном корпусе в 1 млн слов, сравнимом по размеру с Брауновским корпусом [15], был создан словарь под редакцией Засориной для определения базового словаря русского языка [6], который был использован как основа для создания учебного словаря РКИ [10].

В начале XXI в. для русского языка появились большие корпуса, что привело к относительно простому способу создания частотных списков, сначала на основе больших библиотек Интернета[1], потом в рамках проекта Национального корпуса русского языка [5], а затем с помощью специальных программ кролинга для сбора страниц Интернета [20]. Последний корпус был также использован при создании полноценного частотного словаря для преподавания РКИ [23]: в нем выделены значения слов, даны примеры их употребления и перевод (на английский язык).

В данной статье приводятся примеры из частотных списков следующих корпусов:

- Национальный корпус русского языка (НКРЯ), большой представительный корпус, созданный в 2000-е гг. в рамках проекта РАН [5], 90 млн слов в версии частотного списка 2009 г.
- Интернет-корпуса Araneum [8], rutenten [13], ruWac [20], созданные в период 2010-х гг. путем кролинга всех видов текстов на русском языке, каждый из них содержит несколько миллиардов слов.
- Генеральный интернет-корпус русского языка (ГИКРЯ), корпус социальных сетей и средств массовой коммуникации, созданный в 2013–2015 гг. в рамках сотрудничества РГГУ, МФТИ и компании ABBYY [2], в общей сложности 20 млрд слов.

Все частотные списки порождаются путем подсчета числа встречаемости для каждого объекта во всем объеме корпуса. Чаще всего таким объектом подсчета является слово в его словарной форме, но могут считаться и конкретные словоформы, и части речи, и грамматические признаки, и словарные гнезда: например, *учить* и производные от него слова *учитель, учиться, научить, учительский* могут выступать как один объект подсчета. В результате получается ранжированный частотный список, из которого отбираются первые N объектов. Для целей преподавания РКИ традиционным порогом служит ограничение в 5–7 тыс. словарных слов.

Надежность частотных данных

Частотные списки полезны, поскольку человеческая интуиция часто ошибается в определении того, что является более или менее типичным. Например, по замечанию В.И. Беликова, современные толковые словари русского языка снабжают слова *кузен* и *кузина* пометой *устар.*, хотя данные частотных списков показывают, что по частоте их употребления в корпусах они сравнимы со словами *ангина* и *экономный* [1]. Похожим образом с точки зрения исследования коллокаций (статистически частотных словосочетаний) те контексты, которые носители языка предлагают как типичные для слова *глубокий*, касаются объектов с измеряемой глубиной: *след* (901), *яма* (828), *море* (149)[2]. В то же время, по данным корпуса ruWac, намного более частотные словосочетания используют *глубокий* в метафорическом значении: *древность* (3428), *ночь* (2856), *смысл* (2471), *вдох* (2270), *знание* (2006), *понимание* (1899), *сон* (1895), *убеждение* (1825).

Тем не менее сырые данные частотных списков также могут включать очевидные ляпы. Хотя увеличение размера корпуса позволяет собирать данные из большего количества контекстов (что повышает статистическую значимость показателей частоты и надежность информации в частотных словарях), системные тематические сдвиги в составе корпусов могут приводить к неожиданным выбросам частоты. Во всех частотных списках для русского языка зона первых 2 тыс. слов соответствует самому базовому ядру языка, но в него иногда попадают слова, которые интуиция преподавателя скорее всего относит за пределы ядра. Например, для корпуса ruWac рекламные слова и компьютерные термины оказываются в числе базовых слов[3]:

R_{817}: *солнце, защита, половина, запись, боль, пункт, никак,* ***бесплатный****, пожалуйста*;

R_{1264}: *столица, доктор, организовать, ремонт,* ***файл****, наблюдать, существо, поколение, станция*;

[1] bokrcorpora.narod.ru/frqlist/frqlist.html.

[2] Числа в скобках обозначают абсолютную частоту в обсуждаемом корпусе, для обозначения позиции слова в частотном списке используется обозначение R_x.

[3] Это касается и английских интернет-корпусов, например, в корпусе ukWac *file* (R_{614}), *server* (R_{1373}).

R_{2045}: *заходить, любитель, успешно, увеличить,* ***сервер,*** *проще, введение, ожидание, улыбаться.*

В других корпусах возникают другие тематические сдвиги. В частности, НКРЯ содержит тексты из относительно небольшого количества источников, часть слов из которых (например, *театр* из мемуаров, *штамм* из медицинских текстов или *арбитражный* из юридических текстов) попадают, по версии словаря [5], на существенно более высокие позиции в частотных списках в сравнении с интернет-корпусами:

R_{330}: *кто-то, президент, комната, порядок, момент,* ***театр****, читать, письмо, следующий, утро*;

R_{2169}: *смешной, сомневаться, усмехнуться, напротив, несчастный,* ***арбитражный****, случайный*;

R_{3070}: *научить, пациент, сок, сопровождать, тяжесть,* ***штамм****, замена, искусственный.*

В свою очередь, слова повседневной жизни в корпусах обычно представлены меньше, несмотря на их полезность студентам элементарного и базового уровней. Базовый уровень CEFR предполагает умение понимать отдельные предложения и часто встречающиеся выражения, связанные с основными сферами жизни [3]. Опыт построения частотных списков на основе представительных корпусов показывает, что при полностью последовательном проведении принципа частотности многие слова, относящиеся к еде и предметам бытового обихода, оказываются за пределами базового словаря в 5–7 тыс. слов, в отличие от слов *арбитражный* и *сервер*. В Таблице 1 приводятся позиции в частотных списках нескольких слов из этих категорий, полужирным шрифтом выделены слова, которые все-таки попадают в частотный минимум 5 тыс. слов соответствующего корпуса.

Для частичного преодоления этой проблемы тот словник, который был положен в основу учебного словаря [23], был дополнительно обогащен в результате сравнения с переведенными частотными списками с английского, арабского, итальянского и др. языков. Слова русского списка, находившиеся в ruWac ниже порога в 5 тыс. слов, были включены в словник, если они присутствовали в нескольких переводах[4].

Проблема с темами, которые важны для студентов, но редко упоминаются в письменных текстах, связана и с другой проблемой, которая заключается в сильном варьировании частоты слова, поскольку некоторые слова сильнее зависят от темы документа, чем другие. Кен Черч приводит пример со словом *Норьега* [11]: частота этого слова почти во всех документах равна нулю, но если в каком-то документе оно встречается один раз, оно будет повторяться в этом документе примерно так же часто, как и частотные служебные слова. В результате такие тематически зависимые слова попадают в базовый частотный список, вытесняя слова, действительно значимые для студентов.

[4] См. словники этого проекта на сайте corpus.leeds.ac.uk/serge/kelly/.

Таблица 1

Позиции обиходных слов в частотных списках

Слово	НКРЯ	ГИКРЯ, LJ	Araneum	rutenten	ruwac	Словник [23]
арбуз	8712	7721	11259	12652	6995	–
варенье	6844	5306	8762	15988	6671	–
дыня	14905	14104	16207	16565	7948	4970
коньяк	**3049**	**3835**	9057	8279	**3903**	3822
мандарин	17891	7980	12808	15184	7997	4983
наволочка	19443	28562	23010	30367	18228	–
персик	13565	10597	14340	15214	7774	4773
пижама	13745	13067	21182	23514	7801	4790
простыня	5307	7539	12376	11429	**4877**	4360
рюмка	3872	9355	14094	13291	**4645**	4290
уборка	6470	**4318**	**3024**	**3734**	5488	4509
черешня	29417	15682	22620	25629	8341	–

Есть несколько способов учета таких тематических сдвигов, которые основаны на анализе распределения частоты каждого слова по всем документам в корпусе. Простой подход заключается в подсчете частоты слова как количества документов, в которых оно встречается. При этом игнорируется, сколько раз слово встречается в каждом документе. Такой подход справляется с множественным употреблением при тематических сдвигах, но приводит к проблемам, если документы корпуса отличаются по своей длине: слова, часто встречающиеся в коротких документах, оказываются чрезмерно «частотными». На это накладывается еще и то обстоятельство, что в корпус легко получить большое количество коротких документов из одного источника, например, из новостной ленты. В результате частотный список такого рода содержит неоправданно большое количество названий должностей, которые могут встречаться только один раз, но во многих документах. В словарях [4] и [23] в качестве статистической меры учета тематических сдвигов был использован коэффициент Жуйяна, который основан на оценке частоты слова с учетом отклонения от стандартной ошибки среднего при рассмотрении частоты по фрагментам корпуса одинакового размера. Второй статистической мерой для ограничения частоты тематических слов служит винсоризация, т.е. метод, когда оценивается диапазон возможных частот и игнорируются выбросы частоты внутри отдельных документов за пределами этого диапазона [19].

Еще один аспект оценки надежности частотных списков для преподавания касается распределения значений слов. Наличие слова в списке лексического минимума указывает на необходимость владеть только частью его значений. В подавляющем большинстве случаев именно частотные значения наиболее полезны студентам. Например, мы можем взять близкие по частотности слова из списка словаря [23]:

R_{1937}: *пациент, статус, масло, прекращать, ставка, вина, юридический, редкий* и оценить частоту значений многозначных слов с помощью самых частотных коллокаций. Так, слово *масло* чаще всего встречается в значении пищи (самые частотные коллокации *растительное* и *сливочное*), что и должно быть известно студентам. Интересно распределение значений слова *ставка*. Финансовые значения (*процентная, налоговая, тарифная*) и метафорическое (*ставка на развитие, победу, технологию*) на материале ruWac доминируют над более редкими, реализующимися в словосочетаниях *ставка командования, очная ставка, ставка* в азартных играх. В то же время слово *редкий* представляет собой другой случай: оно имеет широкий набор значений, относящихся к регулярной полисемии малого количества (*случаи, виды, гость, волосы*). Если студент понимает общее значение ‘малое количество’ для *редкий*, то он сможет понять все частотные случаи его употребления.

Вариативность языка

Частотные списки отражают среднюю частоту по всему корпусу. Слова, которые часто встречаются в научных текстах, и слова художественной литературы смешиваются в частотном списке в пропорции, в которой соответствующие тексты представлены в корпусе. Данные из одних видов источников, например из новостных лент, можно легко получать в больших объемах, чем из других, в частности из учебно-методических текстов. В результате лексические и грамматические признаки, характерные для источников первого вида, доминируют и не позволяют остальным признакам попасть в ядро языка по частотным спискам. Например, из новостных лент ГИКРЯ в частотный список базового уровня попадают такие слова, как *магнитуда* (R_{3116}) или *полпред* (R_{3466}), но в нем отсутствуют такие слова, как *больно* (R_{14806}) или *давайте* (R_{7365}). В некоторых корпусах, например в ГИКРЯ, проводится сознательная политика отделения списков из разных источников с целью получения более надежной оценки отдельно по каждому компоненту, такому как новостные ленты, социальные сети или художественная литература.

Еще одна проблема связана с жанровой вариативностью внутри самих источников. Байбер заметил, что «вариативность языка между жанрами может быть более значимой, чем между языками» («language may vary across genres even more markedly than across languages») [9]. Если мы берем статистику из корпуса социальных сетей из ГИКРЯ, она смешивает информацию о словах и конструкциях, частотных в личных дневниках, дискуссионных форумах и рецензиях, в трех самых частотных видах текстов социальных сетей. Этот пример показывает также важность лингвистически размеченных корпусов. Например, в корпусе ruWac было проведено автоматическое выделение коммуникативных функций на основе модели функциональных размерностей [18], таких как:

аргумент – дискуссионные блоги, газетные статьи или авторские колонки[5];

инструкт – ЧаВО (FAQs), советы, рецепты или руководства пользователя;

[5] Для каждой функции здесь представлены прототипические примеры. Полный список коммуникативных функций для текстов Интернета с соответствующими определениями обсуждается в [18].

личная – дневниковые записи, блоги путешествий;

новости – лента новостей;

реклама – рекламные тексты и анонсы, спам, объявления о продаже;

худож. – романы, поэзия, мифы, краткое изложение сюжетов.

Таблица 2 показывает некоторые лингвистические признаки, такие как статистика текста, части речи и надежно определяемые синтаксические характеристики, и то, как они коррелируют с соответствующими коммуникативными функциями. Отсутствие знаков + и означает отсутствие статистически значимой корреляции, их количество показывает степень ее наличия в соответствующих текстах.

Количество признаков подсчитывается путем автоматического частеречного разбора [22] или анализом морфемного состава, например, выделением абстрактных существительных по концам -ние, -сть, -изм.

Таблица 2

Коммуникативные функции и распределение лингвистических признаков

Признаки	**аргумент**	**инструкт**	**личная**	**новости**	**реклама**	**худож.**
Длина слова			–	+++	++	–––
Лекс. разнообразие		–	–––	++	++	++
Глаголы, наст.			+			++
Глаголы, прош.		––	++	++	––	+++
Местоимения, 1л.	––	–––	+++	–––		–
Местоимения, 2л.		++		––	++	+
Местоимения, 3л.			––	––	––	+++
Местоимения, вопроситель-ные	++		–			

Предварительный анализ распределения признаков показывает, что рекламные тексты часто оказываются похожими на более формальные тексты с более высокой долей атрибутивных прилагательных, придаточных предложений и длиной слов. С другой стороны, хотя тексты художественной литературы часто используются в качестве примера повседневного общения, по своим признакам, в частности по распределению местоимений и повышенному лексическому разнообразию (TypeToken Ratio), они отличаются от текстов личных дневников, которые служат более надежным приближением к языку повседневного общения.

Корпусные данные такого вида также позволяют помочь преподавателю в оценке работ студентов, в частности в выработке формальных критериев по владению языковыми конструкциями на том или ином уровне владения языком (например, включением требования по успешному использованию отрицаний, придаточных предложений и абстрактных существительных при написании аргументативных эссе).

Подбор иллюстративных примеров

Как мы отметили выше, человеческая интуиция не всегда надежна в отношении оценки частот. Это относится и к подбору иллюстративных примеров: подбор примеров из корпуса позволяет показать студентам реальные контексты употребления слова. Для словаря [23] были разработаны статистические механизмы оценки, которые присваивали каждому кандидату в иллюстративные примеры следующие признаки:

- Лексическая сложность, вычисляемая как доля слов за пределами словника этого словаря за исключением слов, определяемых частеречным анализатором как имена собственные, например *Абрикосов, Варвара* или *Смоленск*. Хотя последние и не попадают в список самых частотных слов, но они и не представляют проблемы для студентов, поскольку последние могут оценить их семантический класс в контексте приведенного примера и его перевода.
- Синтаксическая сложность, вычисляемая как доля существительных, предлогов и союзов, поскольку их количество коррелирует с более сложными речевыми стилями [21].
- Коллокационная полезность, вычисляемая как коэффициент логарифма правдоподобия [12] для оценки связности ключевого слова с его синтаксическими соседями в предложении; такой пример, как *Можно ли кошек кормить капсулами рыбьего жира?*, менее полезен, чем пример *кормить кашей*, поскольку *кормить капсулами* не является устойчивым словосочетанием, даже если слова *капсула* и *каша* входят в лексический минимум.

При подборе иллюстративных примеров рассматривались предложения с минимальной лексической и синтаксической сложностью и максимальной коллокационной полезностью. Например, из общего числа 68 252 реальных корпусных контекстов употребления слова *кормить* остался список из 30 предложений с низкой лексической и грамматической сложностью и высокой коллокационной полезностью, из которых составители словаря выбрали наиболее подходящий пример, приняв во внимание также и педагогические предпочтения.

Надежные данные корпусов позволяют помочь интуиции преподавателя в выборе слов и конструкций, которые нужны студентам. С другой стороны, неправильно считать, что корпус решает все проблемы и полностью заменяет опыт преподавателя. Состав текстов каждого корпуса приводит к существенным изменениям в списках частотных слов и конструкций. Кроме того, эти списки можно строить разными способами, чтобы избавиться от выбросов или чтобы выделить подмножества документов, соответствующие целям обучения. В результате никакой отдельно взятый корпус не может предоставить единственного достоверного списка того, что нужно преподавать на каждом уровне владения языком. С другой стороны, невозможно и игнорировать тот материал, который корпус предоставляет преподавателю в отношении принятия педагогических решений.

Автор благодарен В.И. Беликову и Н.М. Якубовой за полезные комментарии по содержанию статьи. ■

ЛИТЕРАТУРА

1. Беликов В.И. Бабариха и русские системы свойства // Русская словесность. 2020. № 3.
2. Беликов В.И., Копылов Н.Ю., Селегей В.П., Шаров С.А. Дифференциальная корпусная статистика на основании неавтоматической метатекстовой разметки // Труды Диалога Российской конференции по компьютерной лингвистике. Бекасово, 2014.
3. Ирисханова К.М. Система общеевропейских требований к уровням владения иностранным языком: параметры описания, контроля и оценки // Вестник Московского государственного лингвистического университета. 2012. № 648.
4. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка. М., 2009.
5. Сичинава Д.В. Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.
6. Частотный словарь русского языка / Под ред. Л.Н. Засориной. М., 1977.
7. Штейнфельд Э.А. Частотный словарь современного русского литературного языка. Таллин, 1963.
8. Benko V. Two Years of Aranea: Increasing Counts and Tuning the Pipeline // Proc LREC. Portorož, 2016.
9. Biber D., Conrad S. Register, genre, and style. Cambridge University Press, 2009.
10. Brown N.J. Russian learners' dictionary: 10 000 words in frequency order. London, 1996.
11. Church Kenneth. Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p^2 // Proc COLING. Saarbrücken, 2000.
12. Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence // Computational Linguistics. 1993. Vol. 19. No. 1.
13. Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V. The tenten corpus family // Proc Corpus Linguistics Conference. Lancaster, 2013.
14. Josselson H.H. The Russian word count and frequency analysis of grammatical categories of standard literary Russian. Detroit, 1953.

15. Kučera H., Francis W.N. Computational analysis of presentday American English. Providence, 1967.
16. Laufer B., RavenhorstKalovski G.C. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension // Reading in a foreign language. 2010. Vol. 22. No. 1.
17. Nation I. How large a vocabulary is needed for reading and listening? // Canadian modern language review. 2006. Vol. 63. No. 1.
18. Sharoff S. Functional Text Dimensions for the annotation of Web corpora // Corpora. 2018. Vol. 13. No. 1.
19. Sharoff S. Know thy corpus! Robust methods for digital curation of Web corpora // Proc LREC. Marseilles, 2020.
20. Sharoff S., Goldhahn D., Quasthoff U. Frequency Dictionary: Russian / Ed. by U. Quasthoff, S. Fiedler, E. Hallsteindyttir. Leipzig, 2017. Vol. 9.
21. Sharoff S., Kurella S., Hartley A. Seeking needles in the Web haystack: finding texts suitable for language learners // Proc Teaching and Language Corpora Conference, TaLC 2008. Lisbon, 2008.
22. Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Proc Dialogue, Russian International Conference on Computational Linguistics. Bekasovo, 2011.
23. Sharoff S., Umanskaya E., Wilson J. A frequency dictionary of Russian: core vocabulary for learners. London, 2013.

References

1. Belikov V.I. Babariha i russkie sistemy svojstva // Russkaya slovesnost'. 2020. № 3.
2. Belikov V.I., Kopylov N.Yu., Selegej V.P., Sharov S.A. Differencial'naya korpusnaya statistika na osnovanii neavtomaticheskoj metatekstovoj razmetki // Trudy Dialoga Rossijskoj konferencii po komp'yuternoj lingvistike. Bekasovo, 2014.
3. Iriskhanova K.M. Sistema obshcheevropejskih trebovanij k urovnyam vladeniya inostrannym yazykom: parametry opisaniya, kontrolya i ocenki // Vestnik Moskovskogo gosudarstvennogo lingvisticheskogo universiteta. 2012. № 648.
4. Lyashevskaya O.N., Sharov S.A. Chastotnyj slovar' sovremennogo russkogo yazyka. M., 2009.
5. Sichinava D.V. Nacional'nyj korpus russkogo yazyka: ocherk predystorii // Nacional'nyj korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy. M., 2005.
6. Chastotnyj slovar' russkogo yazyka / Pod red. L.N. Zasorinoj. M., 1977.
7. Shtejnfel'd E.A. Chastotnyj slovar' sovremennogo russkogo literaturnogo yazyka. Tallin, 1963.
8. Benko V. Two Years of Aranea: Increasing Counts and Tuning the Pipeline // Proc LREC. Portorož, 2016.
9. Biber D., Conrad S. Register, genre, and style. Cambridge University Press, 2009.
10. Brown N.J. Russian learners' dictionary: 10 000 words in frequency order. London, 1996.
11. Church Kenneth. Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p^2 // Proc COLING. Saarbrücken, 2000.
12. Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence // Computational Linguistics. 1993. Vol. 19. No. 1.
13. Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V. The tenten corpus family // Proc Corpus Linguistics Conference. Lancaster, 2013.
14. Josselson H.H. The Russian word count and frequency analysis of grammatical categories of standard literary Russian. Detroit, 1953.
15. Kučera H., Francis W.N. Computational analysis of presentday American English. Providence, 1967.
16. Laufer B., RavenhorstKalovski G.C. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension // Reading in a foreign language. 2010. Vol. 22. No. 1.
17. Nation I. How large a vocabulary is needed for reading and listening? // Canadian modern language review. 2006. Vol. 63. No. 1.
18. Sharoff S. Functional Text Dimensions for the annotation of Web corpora // Corpora. 2018. Vol. 13. No. 1.
19. Sharoff S. Know thy corpus! Robust methods for digital curation of Web corpora // Proc LREC. Marseilles, 2020.
20. Sharoff S., Goldhahn D., Quasthoff U. Frequency Dictionary: Russian / Ed. by U. Quasthoff, S. Fiedler, E. Hallsteindyttir. Leipzig, 2017. Vol. 9.
21. Sharoff S., Kurella S., Hartley A. Seeking needles in the Web haystack: finding texts suitable for language learners // Proc Teaching and Language Corpora Conference, TaLC 2008. Lisbon, 2008.
22. Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Proc Dialogue, Russian International Conference on Computational Linguistics. Bekasovo, 2011.
23. Sharoff S., Umanskaya E., Wilson J. A frequency dictionary of Russian: core vocabulary for learners. London, 2013.

S.A. Sharoff
University of Leeds
Leeds, UK

NOT ONLY SIZE MATTERS: ISSUES IN CREATING FREQUENCY DICTIONARIES FROM CORPORA

Robustness of frequency data, frequency bursts, genre diversity, illustrative examples, assessment of difficulty

The paper discusses the issues in creating frequency dictionaries aimed at language teaching, while taking into account such parameters as sources of corpora, actual word frequencies, document length consideration, as well as variation in topics and genres. It provides examples of problems with frequency lists and gives recommendations for practical use of frequency dictionaries. In addition to the size of the corpus, the frequency dictionaries are influenced by words that are frequent within long documents, since they lead to frequency bursts, as well as by the link between the topics and genres in a corpus to the learning objectives, since corpora from different subject areas and genres can produce radically different frequency profiles.

НОВОСТИ НОВОСТИ НОВОСТИ НОВОСТИ

В 2020 г. Институт Пушкина по традиции стал участником просветительского проекта Департамента образования и науки г. Москвы «Университетские субботы». Проект стартовал в сентябре 2013 г., и с тех пор наш вуз ежегодно знакомит обучающихся и педагогов столичных образовательных учреждений с актуальными вопросами современного гуманитарного знания. Благодаря «Университетским субботам» гости проекта получили возможность совершенствовать знания в области русского языка и литературы, познавать тайны межкультурной коммуникации, соприкасаться с традициями стран и народов, чему способствует уникальная многонациональная среда института.

Институт Пушкина провел в 2020 г. 52 университетские субботы. Они были посвящены развитию навыков практической грамотности и чтения, коммуникации, различным аспектам русского языка, этнографии и фольклору, истории отечественной и зарубежной литературы, образу Москвы в творчестве отечественных писателей.

С марта 2020 г. университетские субботы проходили в онлайн-формате. Но это не повлияло на заинтересованность и активность участников. Еженедельно виртуальную студию Гос. ИРЯ им. А.С. Пушкина посещают от 70 до 120 учащихся, их родители и учителя, которые тоже участвуют в интерактивных лекциях и мастер-классах: задают вопросы, высказывают мнение по обсуждаемым темам, комментируют.

Проект «Университетские субботы» сохраняет установку на интерактивность. Преподаватели используют на занятиях различные формы для активизации аудитории. Популярны элементы викторины, интеллектуальных игр, тестов на ассоциации и т.д. На мероприятиях, посвященных развитию межкультурной компетенции, студенты получают возможность познакомиться и пообщаться с представителями иных культур, из первых уст узнать об их традициях.

Организаторы «Университетских суббот» института надеются, что и дальше проект будет вызывать интерес у московских школьников и педагогов, что, несмотря ни на какие сложности, наши встречи будут проходить регулярно, а тематика лекций и мастер-классов останется разнообразной и актуальной.

И.С. Леонов,
д-р филол. наук, доцент,
руководитель проекта «Университетские субботы»
в Институте Пушкина