УДК 81'373.47
DOI 10.25205/1818-7935-2019-17-4-63-77

Лексический минимум по языку специальности: сколько слов достаточно? Разработка принципов минимизации

Е. А. Власова, **Е. Л. Карпова**, **М. Ю. Ольшевская**

Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия

Аннотация

Статья содержит исследование методологии составления лексических минимумов (далее – ЛМ) по русскому языку общего владения (Государственный стандарт, Система лексических минимумов В. В. Морковкина и Частотный словарь русского языка для иностранцев, созданный С. А. Шаровым в рамках проекта KELLY), а также анализ специальных лексических минимумов по медицине, робототехнике, ядерной энергетике, математике. Рассмотрены и систематизированы общеметодический, лингвостатистический и корпусно-ориентированный подходы к созданию ЛМ. Статья также описывает процесс и результаты создания собственного корпуса учебников по политологии и частотного списка на основе использования указанных методов. Выявлено, что ключевое место в разработке лексического минимума занимает сам процесс минимизации списка, т. е. принципы, на основании которых определяется длина лексического перечня, а также критерии включения или исключения лексической единицы из финального перечня. Показано, что при корпусно-ориентированном подходе к составлению лексического минимума по отдельной дисциплине ключевую роль играют не только абсолютные и относительные частоты, но и покрываемость – индекс, показывающий, сколько процентов текста составляют все употребления каждой лексической единицы. При составлении лексического минимума по политологии показатель покрытия выявил, что 8 237 самых частотных лексем образуют 98 % токенов всего корпуса. Анализ финального перечня с точки зрения лингводидактики показал, что для понимания 50 % корпуса учебно-профессиональных текстов по дисциплине необходимо знание 1 000 самых частотных знаменательных слов, а оптимальным количеством является интервал в 3 500–4 500 слов частотного списка, после освоения которых рост показателя покрытия замедляется, а понимание обеспечивается в большой степени средствами связности, логической структурой и стратегиями коммуникативного развертывания текста. Результаты показывают, что сочетание трех основных методов создания лексических минимумов позволяет получить методически обоснованное представление о требованиях к словарному запасу студентов, изучающих русский язык в учебно-профессиональных целях.

Ключевые слова

русский язык в специальных целях, лексический минимум, языковые минимумы, корпусная лингвистика, русский язык, критерии отбора лексического минимума, частотность, покрытие

Для цитирования

Власова Е. А., *Карпова Е. Л.*, *Ольшевская М. Ю.* Лексический минимум по языку специальности: сколько слов достаточно? Разработка принципов минимизации // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 4. С. 63–77. DOI 10.25205/ 1818-7935-2019-17-4-63-77

Vocabulary: How Many Words Are Enough? Principles of Minimizing Learners' Vocabulary

Ekaterina Al. Vlasova, **Elizaveta L. Karpova**, **M. Yu. Olshevskaya**

National Research University "Higher School of Economics"
Moscow, *Russian Federation*

Abstract

This article analyses methodology of compiling Russian general wordlists and lexical minima for teaching Russian for specific purposes. The study systematizes three approaches: linguo-didactic, linguo-statistical, and corpus-based. The article also describes the process and results of applying all the three methods to the development of a lexical minimum based on political science corpus. Methodological analysis comprises general word lists for the Russian State Standard Exam (TRKI), the System of lexical minima by V. V. Morkovkin, and the Frequency dictionary of the Russian language for foreigners, created by S. A. Sharov as a part of the KELLY project, as well as special lexical minima for medicine, robotics, nuclear energy, and mathematics. It has been revealed that the core element in the development of a discipline-specific lexical minimum is minimization that involves a set of principles determining the optimal length of the list and lexeme selection. For the Russian general word lists, the most common principles of minimization are methodical expediency ("relevance" of the word at each level), quantitative metrics, including absolute and relative frequencies, the word rank, and a coverage index, showing the percentage of text that every lexeme covers. The article reports the results of combining the quantitative methods, corpus-based analysis, and didactic principles to apply to the development of the lexical minimum based on political science textbooks. The core index, defining the length of this list, was coverage which revealed that 8,237 most frequent lexemes cover 98 % of the whole corpus. The linguo-didactic analysis showed that 1, 000 most frequent lexemes, without stop-words, cover 50 % of this corpus, and therefore this wordlist allows foreign learners to understand about a half of the corpus. After reaching the point of 3,500 of the most frequent words, the coverage index grows insignificantly, and this number can be considered to be a target in teaching and learning discipline-specific vocabulary. It is notable that the recommended lexical minimum, comprising 1,000–3,500 of the most frequent words, is only a starting point for reading comprehension of texts for professionals also referred to as 'special' texts. Their deeper and effective understanding also involves competence in rhetoric strategies and text structure.

Keywords

Russian language for specific purposes, lexical minima, word list, corpus linguistics, Russian language, criteria of lexical selection, word dispersion, word frequency, coverage, number of words

For citation

Vlasova, Ekaterina Al., Karpova, Elizaveta L., Olshevskaya, Maria Yu. Vocabulary: How Many Words Are Enough? Principles of Minimizing Learners' Vocabulary. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2019, vol. 17, no. 4, p. 63–77. DOI 10.25205/1818-7935-2019-17-4-63-77

Вступительные замечания

Русский язык в специальных целях – отдельное направление в лингводидактике со своим кругом методических проблем, которые сформулированы в монографиях [Митрофанова, 1985; Маркина, 2011]. Во-первых, центральное место в преподавании занимает большой объем лексики, необходимой студенту для успешной учебно-профессиональной деятельности, и часто наблюдаемые несоответствия методических материалов реальному содержанию учебных дисциплин. Во-вторых, в отличие от курсов общего владения, на профессионально-ориентированных занятиях отсутствует деление на уровни, а словарный запас студентов в момент поступления в вуз может сильно различаться. В-третьих, профессиональные тексты отражают стилистическое варьирование и многозначность, характерную для узкоспециальных предметных областей.

Для преодоления указанного методического разрыва в отечественной лексикографии активно велась разработка лексических минимумов: под этим термином понимаются как учебные словари, так и комплексные лексические перечни, цель которых – представить сокращенную модель тезауруса, достаточного для осуществления коммуникативной деятельности

на разных уровнях владения языком или в разных предметных областях (см. определение [Маркина, 2011]). С развитием корпусных инструментов решать задачи учебной лексикографии стало проще благодаря возможности автоматически извлекать лексические списки из реальных текстов, используемых в узкоспециальных предметных областях [Sinclair, 2004; McCarthy, 2008]. Однако по мере роста корпусных исследований методисты столкнулись с проблемой, требующей теоретического осмысления [Nation, 2016]: какое количество слов следует считать достаточным для того, чтобы автоматически созданный лексический перечень адекватно отражал значимую лексику предметной области и был методически целесообразным?

Несмотря на растущее число корпусно-ориентированных лексических минимумов, в российской учебной лексикографии указанный вопрос отдельно не обсуждался. Задача данной статьи – проанализировать существующие подходы к минимизации частотного списка и определить оптимальную длину перечня по языку специальности. Исследование проводилось в три этапа:

1) проанализированы корпусно-ориентированные лексические минимумы по языку специальности и выделены методические проблемы, связанные с длиной списков;

2) систематизированы подходы к минимизации списков по языку общего владения и показана их применимость в области профессионально-ориентированных курсов РКИ;

3) на основе выделенных подходов и самостоятельно созданного корпуса учебников по политологии определена оптимальная длина лексического минимума, обеспечивающего понимание большей части текста.

Лексические минимумы по языку специальности: проблемы корпусно-ориентированных подходов

Корпусно-ориентированные исследования, опубликованные в последнее десятилетие, показывают, что рекомендуемая длина лексических минимумов сильно варьируется, а полученные на основе частотных данных списки не всегда полно отражают предметную область. Остановимся на наиболее значимых проблемах, выявленных составителями.

С помощью программы «Wordstat» в МГТУ им. Н. Э. Баумана разработан лексический минимум для специальности «Робототехника» [Ильина, 2013]. Основу корпуса составили все учебные пособия по специальности – всего 26 учебников и учебных пособий объемом 164 523 словоупотреблений, финальный лексический список содержит 1 700 слов. Составители обратили внимание на то, что перечень включал большой процент служебных слов, которые являются самыми частотными, однако для преподавания языка специальности они малосодержательны.

При помощи программы «LitFrequencyMeter» создан частотный список по ядерной энергетике на основе 20 учебников объемом 5 676 страниц. Финальный лексический минимум содержал только 1 000 единиц, а самыми частотными оказались слова, не относящиеся к ядерной тематике [Атийях, 2015. С. 239]:

- общеупотребительная лексика: *это*, *например* и др.;
- общенаучная лексика: *являться*, *следовать*, *позволять*, *определять*, *осуществлять*, *состоять*, *представлять*, *рассматривать*, *использовать*, *зависеть*, *обеспечивать*, *получать*, *требовать*, *применять*, *происходить*, *давать*;
- существительные естественнонаучной направленности: *управление*, *контроль*, *реактор*, *движение*, *безопасность*, *скорость*, *схема*, *пар*, *двигатель*, *генератор*, *энергия* и др.

Как отмечает сам составитель, лингводидактическая проблема полученного списка связана с тем, что он не отражает специфической узкоспециальной лексики по ядерной энергетике.

В 2016–2018 гг. на базе филологического факультета МГУ под руководством О. В. Кукушкиной разработан медицинский лексический минимум [Леоненко и др., 2018], основан-

ный на той же методике, что и английский список общемедицинской лексики MAWL. Исследование включало отбор современных научных статей, различных по тематике – всего 96 статей из 32 областей медицины за 2004–2016 гг. После автоматизации и чистки первичный частотный список составил 15 129 слов, после проверки на омонимию количество лексем в финальном списке составило 10 428 слов за вычетом наиболее частотных служебных слов. При включении / исключении лексем учитывались следующие принципы: словообразовательная активность, абсолютная частотность, регулярность (использование в большом количестве текстов), число родственных слов с данным семантическим элементом [Леоненко и др., 2018]. Исследователи отметили, что в списке, превышающем 10 000 слов, оказалось много лексических единиц, не относящихся к собственно медицинским терминам.

Некоторые корпусные исследования ориентированы только на автоматическое извлечение узкоспециальной терминологии, однако и при такой постановке задачи разброс цифр оказывается значительным. Например, в диссертации [Маркина, 2011. С. 19] описаны внушительные корпуса по философии и педагогике объемом 1 млн и 2 млн словоупотреблений соответственно, однако финальный список содержит всего 100 терминов по каждой предметной области. На базе текстов лекций в МГТУ «СТАНКИН» создан лексический минимум по математике, содержащий около 1 500 слов и сочетаний, а также лексические минимумы по различным специальностям МГТУ МАДИ, включающие 352 слова по математике, 146 слов по физике, 174 слова по химии, 736 слов по биологии [Ильина, 2013].

Представленный выше обзор недавних корпусных профессионально-ориентированных лексических минимумов показывает, что длина перечней варьируется от 1 000−1 700 лексических единиц до 10 000 и более. При этом разные подходы к минимизации приводят к разным методическим проблемам. При небольшой длине списков в 1 000−1 700 единиц список приоритетных лексем содержит много служебных слов, а также общенаучную лексику, при этом узкоспециальная терминология, представляющая трудность для иностранного учащегося, либо не попадает в группу частотных слов, либо извлекается вручную на основе интуитивных представлений. При составлении расширенного списка, содержащего порядка 10 000 лексем, перечень содержит большой объем общеупотребительных и общенаучных слов и не выявляет узкоспециальной терминологии. Кроме того, объем лексического минимума в 10 000 единиц сам по себе представляет трудность для изучения и требует ранжирования и стратификации – деления на темы или уровни. Таким образом, до сих пор не найдено ответов на принципиальные вопросы, связанные с составлением лексического минимума для узкоспециальных предметных областей: нет четкого представления о том, насколько рекомендуемые перечни достаточны для коммуникативной деятельности, не до конца выработаны критерии включения / исключения слов, не обсуждалась возможность их ранжирования. Между тем перечисленные проблемы уже возникали ранее при разработке лексических минимумов общего владения, разработано несколько теоретически обоснованных методов минимизации. В следующем разделе описаны возможности их применения при составлении рекомендуемых лексических списков по языку специальности.

Современные подходы к составлению лексических минимумов по политологии

Лексический минимум в системе тестирования

Наиболее заметное место в современной лингводидактике занимает лексический минимум, положенный в основу Государственного стандарта по русскому языку [Андрюшина и др., 2018]. Эта серия лексических минимумов положена в основу тестирования по русскому языку как иностранному и носит градуальный характер в соответствии с общеевропейской системой тестирования CERF. Система предусматривает 6 уровней владения языком – от A1 до C2, однако российские лексические минимумы разработаны только для 5 уровней – элементарный (A1), базовый (A2), пороговый ТРКИ-3 (B1), необходимый для поступления

в российские вузы, ТРКИ-2 (B2) и ТРКИ-3 (C1). По мнению разработчиков государственного стандарта, уровень C2 приравнивается к уровню носителя и не требует минимизации.

Основные алфавитные лексические перечни перечисленных минимумов создавались на большой лексикографической базе – толковых словарях С. И. Ожегова и Г. Н. Скляревской. Последние издания популярного словаря Ожегова содержат более 100 000 слов, терминов и фразеологических выражений, а «Толковый словарь живого русского языка начала XXI века» Г. Н. Скляревской включает 8 500 слов, в то время как опубликованные лексические минимумы содержат от 3 500 до 10 500 вхождений. Эти данные показывают, что процесс минимизации включал большую работу по исключению лексических единиц. Принимая решение о включении слова в минимум, составители руководствовались общеметодическими критериями [Андрюшина, 2018]: семантическая ценность, способность слова входить в различные сочетания, стилистическая принадлежность, частотность и словообразовательный потенциал слова. Указанные критерии вызывали критику среди исследователей в связи с тем, что решения о включении лексических единиц в основной словарный перечень и соответствии уровню владения языком опираются на субъективный опыт и интуицию составителей [Маркина, 2011].

Длина перечня для каждого уровня в системе ТРКИ основана на методическом представлении о том, что переход на каждый следующий уровень должен сопровождаться удвоением тезауруса (это увеличение мы четко видим в ТРКИ-3,4). По этой причине требования к объему лексического минимума русского языка оказываются на уровнях B2 и C2 выше, чем требования стандартных экзаменов для других языков, ср. объем минимума для кембриджского экзамена по английскому языку [CERF; Meara and Milton, 2003].

Таблица 1

Требования к лексическому минимуму ТРКИ и кембриджского CEF

Table 1

The Comparison of Vocabulary Size for TRFL and CEF

Шкала CERF	ТРКИ	Длина минимума, ед.	Кембриджский экзамен CEF	Длина минимума, ед.
(A1)	Элементарный (ТЭУ)	780	Starters, Movers and Flyers	1 500
(A2)	Базовый (ТБУ)	1 300	Kernel English Test (KET)	1 500–2 500
(B1)	Первый сертификационный уровень (ТРКИ-1)	2 300	Preliminary English Test (PET)	2 750–3 250
(B2)	Второй сертификационный уровень (ТРКИ-2)	5 000	First Certificate in English (FCE)	3 250–3 750
(C1)	Третий сертификационный уровень (ТРКИ-3)	11 000	Cambridge Advanced English (CAE)	3 750–4 500
(C2)	Четвертый сертификационный уровень (ТРКИ-4)	не регламентировано	Cambridge Proficiency in English (CPE)	4 500–5 000

Сравнение данных показывает, что с уровня B2 объем лексического минимума ТРКИ начинает значительно превышать рекомендуемые стандарты кембриджского экзамена: в 1,5 раза на уровне B2, почти в 3 раза на уровне C1. Даже если принять во внимание, что в русском языке видовые пары входят в перечень как две разные лексемы, превышение все равно остается значительным.

Таким образом, система градуального лексического минимума, закрепленного в государственном стандарте, состоит из пяти перечней, максимальный объем словарного запаса, требуемый на уровне С1, составляет около 11 000 лексических единиц. Ключевые критерии минимизации, в том числе решение о включении лексемы в список, порядок следования и соответствие уровню, а также длина списка определялись преимущественно из представлений составителей о порядке изучения русской лексики. Несмотря на критику за субъективный отбор, действующий государственный лексический минимум является на данный момент наиболее объемным и актуальным перечнем из всех существующих.

Система лексических минимумов: лингвостатистический подход

Альтернативу лексическому минимуму, положенному в основу Государственного стандарта, составляет Система лексических минимумов русского языка [Морковкин и др., 1985; 2003]. Дополненное и отредактированное издание [Морковкин и др., 2003] содержит 5 000 слов, разделенных на 10 словарных перечней, которые разработаны на основе масштабного лингвостатистического исследования: решения о включении / исключении слова, длине списка и его делении на уровни принималось на основе количественных показателей.

Лексикографической базой Системы лексических минимумов стали крупные частотные словари русского языка (более подробный обзор см. [Алексеев, 1975]): словарь Джоссельсона 1959 г. (5 230 слов), Частотный словарь Э. А. Штейнфельдт 1963 г. (5 500 слов), словарь Н. П. Вакара 1966 г. (2 380 слов), Словарь разговорной речи 1968 г. (2 380 слов), словарь Л. Н. Засориной 1977 г. (1 024 слов), а также Комплексный частотный словарь русской научной и технической лексики 1978 г. (3 047 слов). Как видно из приведенных данных, при составлении первых русских частотных словарей длина списка варьировалась от 1 000 до 5 500 слов, т. е. разброс цифр меньше, чем в современных корпусных исследованиях.

Система лексических минимумов основана на теоретическом представлении о том, что, создав список слов, зафиксированных в нескольких частотных словарях, основанных на разножанровых источниках, лексикографы могут получить сводный перечень, отражающий лексическое ядро языка [Морковкин и др., 1985]. В основу методологии положено сравнение частотных словарей и данных об употребительности слова, которая измеряется двумя величинами – абсолютной частотой и рангом, т. е. номером в списке, расположенном в порядке убывания частоты. Чем выше частота, тем ближе слово к началу списка и тем ниже его ранг. Чтобы исключить эффект варьирования рангов слова в разных частотных словарях, В. В. Морковкин вводит собственную единую систему ранжирования, поделив список каждого словаря на сотни и присваивая каждому слову индекс в соответствии с номером сотни, в которую оно входит, например, индекс 24 означает, что слово находится в 24-й сотне частотного списка [Там же]. После составления таблицы, содержащей информацию об индексе каждой лексемы в частотных словарях, рассчитана статистическая ценность (степень употребительности) каждого слова – среднее арифметическое его индексов. Полученная таблица упорядочена по степени убывания среднего показателя и снова ранжирована по сотням. В результате анализа выделены слова с подтвержденной высокой употребительностью – это все лексемы, обозначенные индексами от 1 до 35 и встретившиеся среди 2 500 употребительных слов не менее чем в двух частотных словарях [Там же]. Лексические единицы с индексом 35–60, обнаруженные только в одном из частотных словарей среди первых 2 500 слов, выделены в группу слов с неподтвержденной высокой употребительностью [Там же]. Таким образом, ключевыми факторами попадания слова в лексический минимум были его присутствие в нескольких частотных словарях и низкий индекс, то есть близость к вершине списка.

Деление на уровни также проводилось на основе квантитативных метрик: чем меньше индекс слова, тем оно более употребительно и тем ранее следует его изучить студенту. Финальный перечень системы лексических минимумов, состоящий из 5000 слов, разбит на 10 групп по 500 слов в каждой.

При определении финальной длины лексического списка В. В. Морковкин также руководствовался количественными показателями, взяв за основу процент покрытия [Алексеев, 1975] – коэффициент, показывающий, какой процент сумма словоформ одной лексемы или списка составляет от общего числа словоупотреблений. Согласно исследованиям по лингвостатистике, в письменной литературной речи первая 1 000 слов частотного списка покрывает 70–80 % текста [Там же]. Согласно данным В. В. Морковкина, 3 500 лексем из Лексического минимума 1985 г. покрывали 82 % текста [Морковкин и др., 1985]: информация о том, как меняется процент покрытия в зависимости от длины списка, представлена в табл. 2.

Таблица 2

Покрытие системы лексических минимумов

Table 2

Coverage of Lexical Minima

Номер градуального минимума	Длина минимума, ед.	Доля покрытия, %
1	500	58
2	1 000	67
3	1 500	72
4	2 000	76
5	2 500	79
6	3 000	81
7	3 500	83

Предположим, что иностранный студент пассивно владеет 3 500 самых частотных слов, это обеспечит ему понимание 83 % текста. Несмотря на то, что речь идет об идеализированной модели лексического запаса студента, указанные цифры могут служить объективным ориентиром, определяющим длину лексического перечня, достаточного для понимания. Таким образом, сторонники лингвостатистического подхода определяют лексический минимум как список 3 500–5 000 самых частотных слов. Важно учитывать, что составители Системы лексических минимумов использовали консолидацию списка: видовые пары считались как одна лексема, сравнительные и превосходные формы прилагательных и наречий при подсчетах возводились к положительной степени [Морковкин и др., 1985]. Если считать каждую видовую форму отдельным словом, как это делается в других учебных словарях, то рекомендуемый размер минимума необходимо увеличить до 5 000 и выше. Указанный объем лексического запаса в системе ТРКИ, использующей неконсолидированный список, соответствует уровням B2–C1.

Существенное достижение Системы лексических минимумов состоит в разработке объективных, количественно обоснованных критериях минимизации и возможности их применения в современных корпусных исследованиях. Между тем следует принимать во внимание, что лексикографическая база указанного издания основана на частотных словарях начала XX в. и отстает от современного состояния русского языка.

Корпусно-ориентированные учебные словари

Система лексических минимумов В. В. Морковкина опубликована в 2003 г., а уже 27 апреля 2004 г. появился сайт Национального корпуса русского языка (http://ruscorpora.ru), ставший основой для корпусной лингвистики и квантитативного анализа текста на материале русского языка. В настоящий момент объем корпуса достигает более 600 млн токенов.

Преимущества корпусной лингвистики и компьютерных инструментов заключаются в возможности быстрого обследования реального и актуального языкового материала большого объема, а также в применении более сложных статистических расчетов. Одной из первых задач, с которой успешно справилась корпусная лингвистика, было обновление частотных словарей русского языка на материале Национального корпуса русского языка. В настоящий момент в открытом онлайн-доступе находятся материалы Частотного словаря русского языка О. Н. Ляшевской и С. А. Шарова, при составлении которого методология создания частотного перечня была скорректирована с учетом достижений компьютерной лингвистики. В отличие от более ранних частотных словарей, использовавших абсолютные частоты и ранги, основной квантитативной метрикой в корпусных исследованиях по лексикографии являются относительные частоты: как правило, это коэффициент ipm (instances per million words) – абсолютные частоты, поделенные на объем корпуса и умноженные на миллион [Gries, 2015; Ляшевская, 2016]. Использование нормализованных частот позволяет сравнивать корпуса разного объема и сопоставлять полученные на их основе списки с использованием статистических методов, например, хи-квадрата или коэффициента логарифмического правдоподобия. Цель указанных статистических тестов – выявление значимой лексики, специфической для текстов определенного жанра. В частности, на основе Национального корпуса созданы частотные списки значимой лексики художественной литературы, публицистики, непубличистических текстов и разговорной речи.

На основе корпусных методов и статистических вычислений, а также сопоставлении лексических минимумов других языков создан градуальный, то есть поделенный на уровни A1–C2, учебный частотный список по русскому языку как иностранному [Kilgarriff et al., 2014]. Особенность этого лексического минимума состоит в том, что он составлен методом перевода и сведения лексических минимумов нескольких европейских языков: этот список максимально стандартизирует и унифицирует рекомендуемые лексические списки в рамках 6-уровневой системы CERF. Учебный словарь С. А. Шарова также включает частотную лингвоспецифическую лексику и учитывает грамматические свойства русского языка – наличие видовых пар. Соотносительные по виду лексемы посчитаны как отдельные слова. Полученный таким образом сводный учебный лексический минимум A1–C2 содержит 9 000 слов: данная цифра сопоставима с уровнем B2–C1 в системе ТРКИ и Системой лексических минимумов.

Таким образом, разработанные принципы сведения списков, современные корпусные инструменты и статистические методы позволяют решать разнообразные методические задачи и составлять частотные списки лексических единиц для обучения студентов лексике с учетом стилистического варьирования.

Автоматизированное составление лексического минимума по политологии

Основу исследования составил самостоятельно собранный и обработанный корпус учебников по политологии, которые используются в российских вузах. Главным требованием, предъявляемым к любому создаваемому корпусу, является репрезентативность, то есть его способность полно, объективно и адекватно отражать лингвистические свойства исследуемой области [Biber, Reppen, 2015; Баранов, 2003. С. 13]. Для выполнения указанного критерия проанализированы выложенные в открытый доступ списки рекомендованной литературы ведущих российских вузов в области политологии, в том числе МГИМО, МГУ им. М. В. Ломоносова, СПбГУ, РУДН, Высшей школы экономики и др. Далее составлен перечень, содержащий шесть непереводных учебников и учебных пособий, входящих в раздел обязательной литературы к курсу и получивших признание в научном сообществе.

Второй важный критерий надежного корпуса – сбалансированность текстов по их тематике и объему. Собранный корпус содержит современные учебники, изданные в разных науч-

ных центрах с 2004 по 2009 г. и отражающие широкий круг тем, описывающих основы политической теории и истории, политический процесс, отношения и технологии, а также регионалистику. Проверка объема текстов выполнена при помощи ресурса VoyantTool, который автоматически рассчитывает объем каждого текста (см. список учебников в порядке возрастания их объема в табл. 3).

Таблица 3

Корпус университетских учебников по политологии

Table 3

The Corpus of University Textbooks on Political Sciences

Автор	Название	Токены, ед.	Доля покрытия, %
Селютин В. И.	Теория и практика политической науки. Воронеж, 2009	82 457	12
Соловьев А. И., Пугачев В. П.	Введение в политологию. М., 2000	91 079	13
Баранов Н. А.	Политические отношения и политический процесс в современной России: Курс лекций. СПб.: БГТУ, 2004	115 466	16
Соловьев А. И.	Политология: Политическая теория, политические технологии: Учебник для студентов вузов. М., 2006	116 145	16
Макарин А. В.	Теория и история политических институтов. СПб., 2008	103 061	14
Туровский Р. Ф.	Политическая регионалистика. М.: Изд-во ГУ-ВШЭ, 2006	207 785	29
		715 993	100

Из табл. 3 видно, что лексический объем ни одного из учебников не превышает 30 % всего корпуса. Размер корпуса изначально составил 715 993 леммы, однако после чистки и лематизации был сокращен до 696 939 токенов: некорректно было распознано 3 % текста. Первичный частотный список, извлеченный при помощи программы «AntConc», содержал 16 552 леммы – указанная цифра значительно превышает объем лексического минимума уровня ТРКИ-III (около 11 000 лемм), соответствующего уровню С1 в европейской системе тестирования, и в 8 превышает лексический минимум ТРКИ-1, который рассматривается как пороговый для поступления в российские вузы. Сравнение показывает, что первичный список нуждается в методически и статистически обоснованной минимизации.

Первое сокращение перечня выполнено в момент проверки и чистки: в отдельную группу слов (всего 1 213 лексем) выделены собственные имена, аббревиатуры (*КПРФ*, *НАТО*, *ЕС* и др.) географические наименования и прилагательные, образованные от них (например, *аварец – аварский*): указанная информация носит энциклопедический характер и соответствующие лексемы целесообразно оформлять в отдельный справочник персоналий и наименований. После устранения имен собственных и их производных список содержал 15 321 лемму, которые расположены по убыванию частоты в соответствии с законом Ципфа: чем больше порядковый номер слова в списке, тем ниже его частотность.

Важно, что консолидация списка произведена на основе лемматизатора «MyStem»: видовые пары посчитаны как разные лексемы, регулярные формы сравнительной степени и краткие прилагательные считались на основании начальной формы как одна лексема.

Следующий этап обработки списка состоял из исключения слов, встретившихся в учебных текстах по 1–2 раза, то есть представляющих собой феномен hapax legomena [Scott, 2006. С. 26–29]. После удаления из словарного списка лексем с абсолютной частотой < 3 длина полученного перечня составила 8 430 слов: указанный размер словника сопоставим с лексическими минимумами общего владения, в том числе с учебным словарем С. А. Шарова (около 9 000 лексем) и лексическим минимумом ТРКИ-III (11 000 лексем). В методических целях Сокращенный список длиной 8 430 слов был разделен на 2 части: в отдельную группу объединены знаменательные слова, выполняющие номинативную функцию (8 237 лексем), и служебные слова, знакомые иностранным студентам из курсов по русской грамматике (193 лексемы).

Таким образом, в результате проведенного анализа рабочий список, состоявший из 16 534 лексем, разделен на несколько групп: основной перечень, служебные слова, редкие лексемы hapax legomena, имена собственные. В табл. 4 для каждой группы указан суммарный коэффициент покрытия и соотношение объема.

Таблица 4

Частотные списки, составленные на основе корпуса по политологии

Table 4

The Wordlist Strata of the Political Science Corpus

Список	Количество		Покрытие, %
	лексем	словоформ	
Основной перечень	8 237	482 209	69
Служебные слова	193	198 097	29
Hapax legomena	6 891	8 904	1
Имена собственные	1 231	7 729	1
Корпус, всего	16 552	696 939	100

Данные, приведенные в табл. 4, показывают, что служебные слова обладают высоким коэффициентом покрытия, между тем имена собственные и редкие слова (т. е. hapax legomena) традиционно убираются из анализа. После стратификации рабочего списка проведена проверка того, какой процент всего корпуса, состоящего из 696 939 словоформ, образуют самые частотные знаменательные слова, включенные в основной перечень (8 237 лексем). Коэффициент покрытия рассчитывался с шагом в 500 слов в порядке убывания их частоты (табл. 5).

Таблица 5 показывает, что 1 000 самых частотных знаменательных слов финального списка покрывает около 50 % всего корпуса, 7 500 знаменательных слов образуют 69 % словоформ корпуса.

Полученные данные позволяют уточнить предложенную В. В. Морковкиным методическую концепцию, согласно которой первые 1 000–1 500 самых частотных лексем обеспечивают понимание около 60–70 % текста. Данное представление не предусматривает того, что среди самых частотных слов большой процент образуют незнаменательные части речи, отвечающие за грамматическую связность [Пумпянский, 1981. С. 318] и не выполняющие номинативных функций. Данные, приведенные в табл. 5, основаны на предположении о том, что понимание текста обеспечивается в первую очередь знанием знаменательных слов, а не служебных. Согласно информации о покрытии частотного списка, составленного на основе корпуса по политологии, для понимания 50–60 % исследованных учебников требуется знать 1 000–2 000 самых частотных номинативных лексем.

Таблица 5

Покрытие основного частотного перечня

Table 5

The Coverage of the Main Frequency Wordlist

Частотность лексемы	Количество словоформ, ед.	Покрытие, %
1 000	366 002	53
1 500	399 437	57
2 000	420 736	60
2 500	435 278	62
3 000	445 702	64
3 500	453 525	65
4 000	459 589	66
4 500	464 379	67
5 000	468 220	67
5 500	471 431	68
6 000	474 116	68
6 500	476 394	68
7 000	478 394	69
7 500	480 001	69
8 000	481 501	69
8 237	482 209	69

Количественные данные показывают, что рост словарного запаса ведет к увеличению показателя покрытия, а значит и пониманию большего объема текста, до границы в 4 500 лексем. После этой цифры расширение вокабуляра на 500 единиц (с 4 500 до 5 000 лексем) не меняет объема покрытия текста: в рамках этой модели студент, пассивно знающий 4 500 лексем, и студент, знающий 5 000 лексем, будут понимать одинаковое количество текста – 67 %. Еще слабее становится корреляция после порога частотности в 5 500 слов и выше. Таким образом, оптимальной длиной лексического минимума, который бы обеспечивал понимание большей части учебных текстов корпуса, можно считать 4 500 знаменательных слов.

Выводы

Современные компьютерные технологии позволяют быстро создавать корпуса на основе списков рекомендованной литературы и извлекать частотные лексические перечни. Между тем корпусно-ориентированные исследования не дают преподавателю четкой рекомендации по оптимальному объему словника, необходимого для чтения и понимания учебных текстов. Сложилась парадоксальная ситуация: преподаватель может легко узнать, какие лексические единицы являются самыми частотными в профессиональных текстах и в каком порядке их следует изучать, однако не имеет представления об объеме словника, оптимального для усвоения дисциплины. Обзор подходов к составлению лексических минимумов показал, что в качестве объективной величины, отражающей достаточный объем профессионально-ориентированного перечня, может быть использован показатель покрытия, распространенный в прикладной лингвистике, но, на наш взгляд, незаслуженно забытый в корпусно-ориентированных исследованиях по учебной лексикографии.

Наше исследование по созданию лексического минимума на основе рекомендуемых учебников по политологии основано на анализе и совмещении трех методов:

а) общеметодического, основанного на градуальном принципе возрастания сложности;

б) лингвостатистического, основанного на показателе частотности и коэффициенте покрытия лексемы, – показателе, который отражает процент употреблений слова от общего количества словоформ в тексте;

в) корпусно-ориентированного, основанного на получении частотного списка и его сопоставлении с другими списками для выделения ядра лексикона и стилистической стратификации.

На основе существующих подходов разработан комплексный метод минимизации первичного частотного списка, исключающий собственные имена, редко используемые лексемы и служебные слова, не выполняющие номинативных функций. В результате применения данных критериев частотный список сокращен с 16 552 до 8 237 лексем. Для финального перечня знаменательных слов рассчитан коэффициент покрытия с шагом в 500 слов.

Анализ финального перечня показал, что 1 000 самых частотных знаменательных слов покрывают 50 % корпуса по политологии. После границы в 3 500 самых частотных знаменательных слов увеличение коэффициента покрытия замедляется. С точки зрения лингводидактики это можно интерпретировать так: после достижения объема словарного запаса, состоящего из 3 500 самых частотных слов, изучение 1 000 новых лексических единиц не приводит к значительному улучшению понимания – на этом этапе большее значение имеют средства связности, организация и логика развития текста рассуждения или текста-описания. Следовательно, при создании корпусно-ориентированных списков для обучения языку специальности минимальным объемом можно считать 1 000 знаменательных слов, покрывающих около 50 % корпуса. Использованный метод показывает, что оптимальным количеством является объем в 3 500–4 500 лексических единиц. Отметим, что указанные цифры не учитывают служебные слова: они исключены из списка в связи с тем, что в методике преподавания русского языка как иностранного служебные части речи является частью изучения грамматики.

Список литературы

Андрюшина Н. П. Лексический минимум по русскому языку как иностранному. Третий сертификационный уровень. Общее владение. СПб.: Златоуст, 2018.

Атийях Э. А. Профессиональный лексический минимум по ядерной энергетике для иностранных магистрантов // Инновации и инвестиции. 2015. № 8. URL: https://readera.ru/142163867 (дата обращения 22.06.2019).

Алексеев П. М. Статистическая лексикография. Л., 1975. 120 с.

Баранов А. Н. Введение в прикладную лингвистику. М.: Эдиториал УРСС, 2001. 360 с.

Ильина О. А. Лексический минимум по языку специальности «Робототехника» как основа формирования лингвокоммуникативной компетенции иностранных магистрантов // Гуманитарный вестник. 2013. Вып. 2 (4). URL: http://hmbul.bmstu.ru/catalog/lang/ling/42.html (дата обращения 22.06.19).

Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. М.: Рукописные памятники Древней Руси, 2016. 518 с.

Ляшевская О. Н., **Шаров С. А.** Электронная версия издания: Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. URL: http://www.artint.ru/projects/frqlist.php (дата обращения 22.06. 2019).

Леоненко А. Д., **Шерварлы М. Г.**, **Ямилова Д. А.**, **Варивода Н. С.** Проблема отбора общепрофессиональной лексики для обучения иноязычных студентов (устный) // Метапредметный подход в образовании: Русский язык в школьном и вузовском обучении разным предметам: Сб. ст. Межрегион. науч.-практ. конф. / Сост. О. Е. Дроздова. М.: МПГУ, 2018. 372 с.

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.

Митрофанова О. Д. Научный стиль речи: проблемы обучения. 2-е изд., перераб. и доп. М.: Рус. яз., 1985. 128 с.

Морковкин В. В, **Сафьян Ю. А.**, **Степанова Е. М.**, **Дорофеева И. В.** Лексические минимумы современного русского языка. М.: Рус. яз., 1985. 609 с.

Морковкин В. В., **Богачева Г. Ф.**, **Луцкая Н. М.**, **Попова З. П.** Система лексических минимумов современного русского языка: 10 лексических списков от 500 до 5000 самых важных русских слов / Под ред. В. В. Морковкина. М.: АСТ: Астрель, 2003.

CERF – Общеевропейские компетенции владения иностранным языком: изучение, обучение, оценка. Страсбург: Департамент по языковой политике, 2003. 256 с.

Пумпянский А. Л. Введение в практику перевода научной и технической литературы на английский язык. 2-е изд., доп. М.: Наука, 1981. 344 с.

Савина О. Ю. Методика формирования лексического минимума с помощью конкордансера // Вестник Тюмен. гос. ун-та. Гуманитарные исследования. Humanitates. 2016. Т. 2, № 1. С. 92–99.

Gries, **S.** Statistics for learner corpus research. In: S. Granger, G. Gilquin, F. Meunier (eds.). The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics. Cambridge, Cambridge University Press, 2015.

Biber, **D.**, **Reppen R**. (eds.). Handbook of Corpus Linguistics. Cambridge, Cambridge University Press, 2015.

Kilgarriff, **F.** Charalabopoulou, M. Gavrilidou et al. Corpus-based vocabulary lists for language learners for nine languages. In: Lang Resources & Evaluation, 2014.

Meara P., **Milton J.** X_Lex, The Swansea levels test. Newbury, Express, 2003.

McCarthy, **M.** Accessing and interpreting corpus information in the teacher education context. *Language Teaching*, 2008, vol. 41.

Nation, **I. S. P.** Making and Using Word Lists for Language Learning and Testing. John Benjamins, 2016.

Sinclair, **John McH.** How to use corpora in language teaching. Amsterdam, Benjamins, 2004, 307 p.

Scott, M., **Tribble, C.** Textual Patterns: Key words and corpus analysis in language education. Amsterdam, Benjamins, 2006, 200 p.

Список источников

Баранов Н. А. Баранов. Политические отношения и политический процесс в современной России: Курс лекций. СПб.: БГТУ, 2004.

Макарин А. В. Теория и история политических институтов: Учеб. пособие для вузов / Под ред. А. В. Макарина, А. И. Стребкова. СПб., 2008.

Селютин В. И. Теория и практика политической науки. Воронеж, 2009.

Соловьев А. И., **Пугачев В. П.** Введение в политологию. М., 2000.

Соловьев А. И. Политология: Политическая теория, политические технологии: Учебник для студентов вузов. М., 2006.

Туровский Р.Ф. Политическая регионалистика. М.: Изд-во ГУ-ВШЭ, 2006.

References

Alekseev, **P. M.** Statistical lexicography. Leningrad, 1975, 120 p. (in Russ.)

Andryushina, N. P. A minimal Russian wordlist for foreign learners. Third certificate level. General knowledge. St. Petersburg, Zlatoust, 2018. (in Russ.)

Attiyah E. A. Professional'nyi lersicheskii minimum po yadernoi energetike dlya inostrannykh magistrantov. *Innovations and Investments*, 2015, no. 8. (in Russ.) URL: https://readera.ru/142163867 (date accessed 22.06.2019).

Baranov, A. N. Introduction to applied linguistics. Moscow, Editorial URSS, 2003, 360 p.

Biber, **D.**, **Reppen R**. (eds.). Handbook of Corpus Linguistics. Cambridge, Cambridge University Press, 2015.

CERF – Common European framework of reference for languages: learning, teaching, assessment. Strasbourg: Department of language policy, 2003, 256 p. (in Russ.)

Gries, **S.** Statistics for learner corpus research. In: S. Granger, G. Gilquin, F. Meunier (eds.). The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics. Cambridge, Cambridge University Press, 2015.

Ilyina, O. A. Lexical minimum of English for special purposes "Robotics" as the basis for the formation of linguo-communicative competence of foreign mA – gitrento. *Humanitarian Bulletin*, 2013, iss. 2 (4). (in Russ.) URL: http://hmbul.bmstu.ru/catalog/lang/ling/42.html (date accessed: 22.06.2019).

Kilgarriff, **F.** Charalabopoulou, M. Gavrilidou et al. Corpus-based vocabulary lists for language learners for nine languages. In: Lang Resources & Evaluation, 2014.

Leonenko, **A. D.**, **Servarly, M. G.**, **Amirova, D. A.**, **Varivoda, N. S.** The problem of selection professional vocabulary for learning foreign language students (oral). In: Metasubject approach in education: Russian language in school and University teaching different subjects: collection of articles of Interregional scientific and practical conference. Comp. by O. E. Drozdova. Moscow, Moscow State University Press, 2018, 372 p. (in Russ.)

Lyashevskaya, O. N. Corpus tools in grammatical studies of the Russian language. Moscow, Handwritten monuments of Ancient Russia, 2016, 518 p. (in Russ.)

Lyashevskaya, O. N., **Sharov, S. A.** Electronic version of the publication: Frequency dictionary of the modern Russian language (based on the materials of the National corpus of the Russian language). Moscow, Azbukovnik, 2009. (in Russ.) URL: http://www.artint.ru/projects/frqlist.php (date accessed: 22.06.19)

Markina, **E. I.** Linguodidactic basis for creating minimal wordlists for Russian foreign learners. Abstract. Moscow, 2011, 24 p. (in Russ.)

McCarthy, **M.** Accessing and interpreting corpus information in the teacher education context. *Language Teaching*, 2008, vol. 41.

Meara P., **Milton J.** X_Lex, The Swansea levels test. Newbury, Express, 2003.

Mitrofanova O. D. Scientific style of speech: teaching issues. 2nd ed., improved. Moscow, Russian Language Publ., 1985, 128 p. (in Russ.)

Morkovkin, V. V., **Bogacheva, G. F.**, **Lutskaya, N. M.**, **Popova, Z. P.** System of lexical minima of the modern Russian language: 10 lexical lists from 500 to 5000 of the most important Russian words. Ed. by V. V. Morkovkin. Moscow, AST: Astrel, 2003. (in Russ.)

Morkovkin, V. V., **Safyan**, **Y. A.**, **Stepanov, E. M.**, **Dorofeeva, V. I.** Lexical minimums of contemporary Russian language. Moscow, Russian Language Publ., 1985, 609 p. (in Russ.).

Nation, **I. S. P.** Making and Using Word Lists for Language Learning and Testing. John Benjamins, 2016.

Pumpyansky, **A. L.** Introduction to English translation of research and technical literature. Moscow, Nauka, 1981, 344 p. (in Russ.)

Savina, **O. Yu.** The Method of Forming a Vocabulary Minimum by Applying Corpus Analysis Toolkit for Concordancing. *Tyumen State University Herald. Humanities Research*, 2016, vol. 2, no. 1, p. 92–99. (in Russ.)

Scott, M., **Tribble, C.** Textual Patterns: Key words and corpus analysis in language education. Amsterdam, Benjamins, 2006, 200 p.

Sinclair, **John McH.** How to use corpora in language teaching. Amsterdam, Benjamins, 2004, 307 p.

Source List

Baranov, N. A. Political relations and political process in modern Russia: Lectures. St. Petersburg, BSTU Press, 2004. (in Russ.)

Makarin, A. V. Theory and history of political institutions: textbook for universities. Eds. A. V. Makarin, A. I. Strebkov. St. Petersburg, 2008. (in Russ.)

Selyutin, V. I. Theory and practice of political science. Voronezh, 2009. (in Russ.)

Solovyov, A. I., Pugachev, V. P. Introduction to political science. Moscow, 2000. (in Russ.)

Soloviev, A. I. Political Science: Political theory, political technologies: Textbook for University students. Moscow, 2006. (in Russ.)

Turovsky, R. F. Political regions. Moscow, Publishing house of HSE, 2006. (in Russ.)

Материал поступил в редколлегию
Date of submission
19.07.2019

Сведения об авторах / Information about the Authors

Власова Екатерина Александровна, кандидат филологических наук, старший преподаватель Школы лингвистики Национального исследовательского университета «Высшая школа экономики» (ул. Мясницкая, 20, Москва, 101000, Россия)

Ekaterina Al. Vlasova, PhD, Senior Lecturer of the Faculty of Humanities, School of Linguistics, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow, 101000, Russian Federation)

evlasova@hse.ru
ORCID 0000-0001-6121-1934
SPIN 9231-4765

Карпова Елизавета Львовна, магистрант факультета гуманитарных наук школы лингвистики Национального исследовательского университета «Высшая школа экономики» (ул. Мясницкая, 20, Москва, 101000, Россия)

Elizaveta L. Karpova, Master Graduate of the Faculty of Humanities, School of Linguistics, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow, 101000, Russian Federation)

lizakarpova95@mail.ru
ORCID 0000-0001-8405-2518
SPIN 8074-4047

Ольшевская Мария Юрьевна, старший преподаватель Школы лингвистики Национального исследовательского университета «Высшая школа экономики» (ул. Мясницкая, 20, Москва, 101000, Россия)

Maria Yu. Olshevskaya, Senior Lecturer of the Faculty of Humanities, School of Linguistics, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow, 101000, Russian Federation)

molshevskaya@hse.ru
ORCID 0000-0002-1050-0784
SPIN 2146-2398