Подходы к составлению лексических минимумов в России и за рубежом: проблемы и перспективы

Н. А. Муравьев, М. Ю. Ольшевская

*Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия*

Аннотация

Представлен сравнительный обзор подходов, которые практиковались в российской традиции составления лексических минимумов русского языка, а также в зарубежных традициях, среди которых особое внимание уделено англо- и немецкоязычной традиции. Цель обзора – проследить направление развития науки в области составления лексических списков на стыке лексикографии и лингводидактики и определить критерии, на которые должны ориентироваться составители списков, чтобы отвечать запросам пользователей в рамках различных задач. Приводится определение понятия «лексический минимум», и задаются параметры сравнения лексических минимумов в последующем обзоре. Рассматриваются отечественные подходы к составлению списков с фокусом на лингводидактической проблематике, лежащей в основе всего русскоязычного направления. Обсуждаются зарубежные подходы, основное внимание уделяется спискам общего пользования и проблеме выделения лексического ядра. Сравниваются отечественная и зарубежные традиции. Приведенный в статье обзор дает основания утверждать, что вне зависимости от предназначения лексического минимума создание его требует сочетания как статистического, так и коммуникативно-ориентированного метода. Кроме того, для составления актуального и достоверного корпуса необходимо равномерное соотношение составных частей анализируемых данных: кроме корпуса текстов художественной литературы следует обратиться к данным устного корпуса и к таким различным по стилю и жанру источникам, как газетный, художественный и академический корпусы, а также корпус интернет-речи.

Ключевые слова

языковые минимумы, список слов, корпусная лингвистика, русский язык, критерии отбора лексического минимума, дисперсия слов, частотность, дисперсия, покрытие

Для цитирования

Муравьев Н. А., Ольшевская М. Ю. Подходы к составлению лексических минимумов в России и за рубежом: проблемы и перспективы // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 1. С. 78–89. DOI 10.25205/1818-7935-2019-17-1-78-89

Approaches to the Composition of Lexical Minima in Russia and Abroad: Problems and Prospects

Nikita A. Muravyev, Maria Yu. Olshevskaya

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

Abstract

This paper presents an overview of the existing approaches in Russia and abroad to the compilation of minimal vocabulary lists. Special attention is paid to the English, as well as to the German scientific tradition. The purpose of the overview is to track and compare the underlying lexicographical and lingvo-didactic trends from the beginning of the 20th century until now and to define the criteria for making a list that would match the expectations of the modern user. The first section of the article provides a definition of the notion *lexical minimum* and introduces the parameters of comparison for the wordlists under discussion. By lexical minimum we understand not only foreign learners' list but more broadly any wordlist which is compiled by minimization of the lexicon on the basis of statistical, pragmatic or mixed criteria. The wordlists are compared across four parameters: purpose of the list (general service list,

theoretical, lingvo-didactic), approach to compilation (statistical, pragmatic, mixed), sources of data (corpora, textbooks, questionnaires etc.) and text coverage in percent. The second section discusses the existing approaches to the lexical minima compilation in Russia with emphasis on pedagogical aspects which are prominent in the Russian tradition. The second section discusses the approaches used in German and English lexicographical traditions focusing on the problem of defining core vocabulary and compiling general service lists. The closing section of the article compares Russian and foreign traditions and summarizes the overview. The present overview suggests that the creation of minimal vocabulary lists requires a combination of both statistical and communicatively oriented methods. In addition to that, given the recent development of large corpora the new challenge arises as to provide a stylistically diverse and balanced corpus or a number of corpora that would serve as a proper vocabulary basis for a vocabulary list. Thus, in addition to the fiction texts, authors should include oral corpus data, as well as newspaper, art and academic sources, and internet speech.

Keywords

lexical minima, word list, corpus linguistics, Russian language, word list criteria, word dispersion, frequency, distribution, lexical overlap

For citation

Muravyev, Nikita A., Olshevskaya, Maria Yu. Approaches to the Composition of Lexical Minima in Russia and Abroad: Problems and Prospects. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2019, vol. 17, no. 1, p. 78–89. (in Russ.) DOI 10.25205/1818-7935-2019-17-1-78-89

Введение

Концепция лексического минимума как набора наиболее значимых лексем языка, используемого в разнообразных практических целях, в том или ином виде существует уже не одно столетие. Подобные списки составлялись для разных теоретических и прикладных областей: образования, лексикографии, стенографии, философии языка, академического письма и многих других. По мере эволюции данной практики неоднократно менялось представление о том, какие лексемы языка являются значимыми и как правильно определять эту значимость в зависимости от поставленной задачи. Одни считают, что объективной мерой значимости является частотность слов в тексте, другие определяют значимость субъективно, но при этом руководствуются практическими соображениями вроде тематической репрезентативности, удобства представления, а третьи придерживаются промежуточных взглядов. Поэтому все многообразие подходов нуждается в осмыслении, сопоставлении и оценке.

Данная статья представляет собой обзор существующих подходов к составлению лексических минимумов в российской и в зарубежных традициях, среди которых особое внимание будет уделено наиболее влиятельным англо- и немецкоязычной традиции. Цель обзора – проследить направление развития науки в области лексических минимумов и определить наиболее значимые критерии для их составления.

Определение лексического минимума

Лексические минимумы имеют различную структуру и состав и создаются с самыми разными целями. Поэтому, для того чтобы иметь возможность сравнивать такие списки между собой, прежде всего, необходимо задать рамки того, что входит в это понятие, и определить параметры сравнения, релевантные для всех рассматриваемых списков. В своем изначальном, узком понимании лексический минимум представляет собой небольшого объема список слов, которые необходимо знать изучающим иностранный язык на определенном уровне владения языком. Чаще всего при этом слова для списка отбираются исходя из субъективных представлений составителей о коммуникативных потребностях изучающих иностранный язык. Такого рода списки составляют основу русскоязычной традиции, хотя в небольшом объеме встречаются и в зарубежных традициях (ср. [Bakonyi, 1934; Thorndike, Lorge, 1944; Carrol et al., 1971]). Параллельно с этим в зарубежной лексикографии возникает и получает развитие отдельная традиция составления списков «общего пользования», отражающих наиболее употребительный пласт лексики, или лексическое ядро языка (ср., к примеру, [West, 1953; Leech et al., 2001; Brezina, Gablasova, 2015]). Эти списки составляются в результате от-

бора, но отбор происходит не на коммуникативной, а на статистической основе с использованием коллекций текстов и полноценных корпусов. Такие списки в строгом смысле нельзя назвать лексическими минимумами, поскольку содержащаяся в них лексика не является минимально необходимой для каких-либо практических целей. Однако в целом, поскольку в эпоху больших данных составление списков без учета статистики фактически теряет смысл, для адекватной оценки текущей ситуации в области составления лексических минимумов исключительно необходимо учитывать и второй вид списков. По этой причине в данной статье предлагается расширенное понимание лексического минимума (далее ЛМ) как *любого списка лексики, полученного в процессе минимизации словарного состава языка ручным или автоматическим методом с учетом коммуникативных, статистических или смешанных критериев*.

Поскольку множество таких списков весьма разнородно, мы будем опираться на перечень параметров, которые будут служить ориентиром для обсуждения и сравнения рассматриваемых ЛМ. В число предлагаемых параметров входят следующие:

- предназначение (общего пользования, теоретические, учебные);
- подход к составлению (коммуникативный, статистический, смешанный);
- источники данных (корпус письменных текстов, материал опроса носителей, прочее);
- процент покрываемости текстов.

Первый параметр различает списки лексики, создаваемые для общего пользования в качестве теоретической модели и в помощь изучающим язык как иностранный. Второй параметр подхода к составлению различает опору на коммуникативные критерии (субъективные, тематические), на статистические критерии (объективные измерения частотности, дисперсии слов и др.) и комбинированные подходы. Следующий параметр характеризует источники данных, на основе которых строится список, а именно: текстовые корпуса, словари, учебная литература, опросники и прочее. Наконец, некоторые составители списков измеряют валидность полученного списка, тестируя степень покрытия их списком произвольного текста. Наличие таких данных также будет учитываться в настоящем обзоре.

Подходы к составлению лексических минимумов в России

Отечественными исследователями с 1950-х гг. предлагалось порядка 18 разных принципов отбора и формирования ЛМ, как, например, принцип стилистической нейтральности слова, сочетаемости, семантической и словообразовательной ценности, частотности и др. [Маркина, 2011]. В 1960–1980-е гг. И. В. Рахманов, И. Д. Салистра, В. Г. Костомаров, Э. А. Штейнфельдт, В. В. Морковкин, П. Н. Денисов разрабатывали и другие критерии отбора лексики. В результате основой для создания существующих ЛМ стали такие критерии, как частотность, сочетаемость, стилистическая нейтральность, семантическая и словообразовательная ценность, многозначность, и русская лексикографическая традиция опиралась на них.

В. В. Морковкин полагал, что основные проблемы при создании ЛМ сводятся к определению количества слов, источника их отбора и реализации процедуры по отбору [Морковкин, Дорогонова, 1976]. Вокруг критериев отбора релевантной лексики создания ЛМ развернулась полемика. Ю. А. Марков считал принцип частотности абсолютным при составлении ЛМ, а В. Г. Костомаров, В. В. Морковкин говорили о недостаточности критерия частотности как ведущего принципа создания ЛМ и необходимости учитывать практическую значимость, потребности пользователя. В. Г. Костомаров указывал, кроме того, что начиная со второй тысячи слов сложно провести границу по частотности, потому что все слова становятся примерно одинаково частотны [1963].

Еще одна из актуальных проблем при выделении ЛМ состоит в том, что слова, необходимые в повседневном общении, не всегда являются частотными. Поэтому один из первых ЛМ, созданный в 1930 г. для английского языка, не содержит таких слов, как «до свидания», «спасибо», «большой», «никогда», «хотеть» [Маркина, 2011]. В. В. Морковкин предлагал

решить эту проблему введением «субъективного» критерия отбора лексики путем проведения опроса среди школьников на предмет выделения самых необходимых слов в повседневном общении [Морковкин, Дорогонова, 1976].

Говоря об определении объема списка, также можно увидеть различие в подходах. В зарубежной лексикографии в связи с понятием коммуникативной компетенции в основу создания «Общеввропейской компетенции владения языком» легла концепция Я. А. Ван Эка [van Ek, 1987; Трушина, 2013]. Согласно данной концепции, объем словаря начального уровня зависит от целей общения на начальном этапе (для английского ЛМ это умение рассказать о себе, семье, ближайшем окружении, понимать вывески, объявления, открытки, заполнять анкеты) и составляет 120 слов (survival list), а для повседневного общения потребуется порядка 2 000 слов.

Сходным образом представлен и круг тем в ЛМ русского языка для элементарного владения [Владимирова и др., 2001]: «Рассказ о себе», «Работа», «Учеба», «Рабочий день», «Свободное время», «Отдых», «Семья». Слова, вошедшие в минимум, дают также возможность общаться на такие темы, необходимые для уровня выживания в иноязычной среде, как «Питание», «Здоровье», «Погода», «Транспорт», «Покупки», «Город, ориентация в городе». Он включает также ограниченный объем лексики, описывающей атрибуты учебного процесса (преподаватель, учебник и др.), которая менее значима для повседневного общения, но является неотъемлемой частью обучения иностранному языку. Кроме того, в минимуме представлены различные формулы речевого этикета, необходимые изучающим русский язык как иностранный при решении актуальных коммуникативных задач.

Кроме того, существуют разногласия по поводу выбора источников текстов для создания ЛМ. Основное противоречие состоит в том, что для создания ЛМ русского языка ранее брались словари частотности (например, [Штейнфельдт, 1963; Засорина, 1977; Лёнгрен, 1993]), толковые и переводные словари, списки лексики к учебникам (см., например, словарь «2 380 слов, наиболее употребительных в русской разговорной речи», составленный коллективом авторов Университета дружбы народов им. П. Лумумбы), а не собственно тексты из художественной литературы, СМИ и т. п. [Морковкин, Дорогонова, 1976].

По мнению Е. И. Маркиной [2011], практика использования НКРЯ до недавнего времени в значительной мере недооценивалась составителями ЛМ. Вместе с тем была произведена попытка создать ЛМ с помощью автоматизированных систем для выделения терминологического словника по методологии А. А. Поликарпова. В результате с помощью морфоанализатора DicTUM, разработанного в Лаборатории компьютерной лексикологии и лексикографии филологического факультета МГУ им. М. В. Ломоносова (О. В. Кукушкина, А. А. Поликарпов), было отобрано по 100 слов для двух специальностей [Маркина, 2011].

Из современных подходов можно отметить совместную работу филологического факультета Барселонского университета и филологического факультета МГУ им. М. В. Ломоносова. Коллективом предпринята попытка создания иллюстративного интерактивного ЛМ. Этот проект разрабатывается на базе существующих лексических минимумов ТРКИ для испанских студентов, изучающих русский язык в рамках специализации «Славянская филология» [Дунаева, Соррилья, 2010]. Ресурс предполагает работу с лексикой на различных уровнях: семантическом, синтаксическом, грамматическом. Для каждого слова в этом ЛМ указывается ударение, произношение, лексическая, толковая, графическая иллюстрации, а также перевод. Ресурс представлен серией заданий, направленных на усвоение обучающимся лексики на основных уровнях, включающих словосочетание и текст. Данный ресурс, несомненно, представляет большой интерес для учебных целей, однако не совершенствует существующий ЛМ, поскольку список слов ресурса базируется на существующих ЛМ русского языка как иностранного.

Говоря о новом типе ЛМ русского языка, В. В. Морковкин предлагал считать «единицей учета при освоении лексики... не слово как таковое, а слово в определенном значении» [Морковкин, Дорогонова, 1976]. Вместе с тем в российских ЛМ в качестве единицы описа-

ния по-прежнему выступает слово в совокупности его значений. В некоторых списках, составленных европейскими авторами (например, в [West, 1953]), единицей описания является не отдельное слово, а так называемая «словарная семья» (= word family – базовое слово во всех его значениях, в том числе в разных частеречных значениях) и его дериваты, которые учащиеся могут понять, опираясь на знание правил словообразования. Тем не менее в силу слишком разной употребительности дериватов одной лексемы и по некоторым другим причинам эта концепция так и не получила широкого распространения.

В. В. Морковкин предлагал также идею создания универсального ЛМ, состоящего из двух частей: 1) списка слов, отобранных на основе частотных словарей, а также на основе «интуиции» составленного тематического «распространителя» алфавитного списка с нумерацией тем и пометой к каждому слову; 2) материалов для поаспектной работы над русской лексикой, «трудностей» русского языка, включающих синонимы, антонимы, паронимы, омонимы, несвободные сочетания и т. д. Слова из первого списка автор предлагал считать обязательными для усвоения, а из второго – факультативными. Такое распределение позволило бы избежать таких пробелов, как наличие в основном корпусе слова «малина» при отсутствии слова «смородина», хотя оба они принадлежат к одной тематической группе [Морковкин, Дорогонова, 1976]. Однако в существующем ЛМ элементарного уровня русского языка [Андрюшина, Козлова, 2006] такое разделение не очевидно: в описании ЛМ нет информации, все ли слова из списка являются одинаково важными и учитываемыми при тестировании элементарного уровня (ТЭУ). Тем не менее, проанализировав данный ЛМ, можно увидеть, что в общих чертах он соответствует критериям В. В. Морковкина. ЛМ состоит из 3 частей, содержит алфавитный список частотных слов, тематические группы, последняя часть издания – приложения, о которых писал В. В. Морковкин.

Таким образом, можно сказать, что в отечественной лексикографии проблема выделения ЛМ до сих пор является актуальной. Среди основных спорных вопросов остаются, на наш взгляд, следующие:

- проблема отбора слов для ЛМ (количество, способ отбора, источники отбора);
- круг предметно-тематических областей («тем»), которые соотносятся со списком важнейших слов для владения русским языком на элементарном (начальном) уровне;
- необходимость отнесения частотности к ведущим критериям отбора слов в ЛМ.

Практика составления ЛМ в России оказывается при этом почти полностью ориентированной на лингводидактическую задачу обучения русскому языку как иностранному. В качестве основных их источников используются преимущественно ранее составленные ЛМ и частотные списки, которые модернизируются и дополняются данными анкетных опросов носителей и другими вспомогательными материалами.

Учитывая, что появление первых теоретических работ по вопросу создания ЛМ в отечественной лексикографии относится к 1950-м гг., а в зарубежной науке – к 1930-м гг., обратимся далее к истории вопроса за рубежом.

Подходы к составлению лексических минимумов за рубежом

Попытки создания ЛМ были предприняты в рассматриваемых европейских странах (Англия, Франция, Германия) уже в 20–30-х гг. XX в.

В 1930 г. Ч. К. Огден создал словарь базовой английской лексики Dictionary of Basic English, состоящий из 850 слов [Ogden, 1930]. Концепция данного словаря основана не на принципе частотности, а на идее поиска минимального лексикона, достаточного для выражения главных идей о мире. Данный ЛМ позиционировался автором как мини-модель языка, способная заменить естественный язык в качестве средства общения. Другими словами, это был искусственный язык, созданный автором из лексических единиц, отобранных для пояснения основной информации о мире.

Впоследствии в словаре было отмечено два существенных недостатка: 1) не разрешается проблема многозначности отобранных слов (было подсчитано, что 850 единиц, вошедших

в минимум Огдена, имеют более 12 000 значений; 2) отсутствует целый пласт слов, необходимых для повседневного общения, но не выражающих каких-либо «идей» о мире (например, «до свидания», «спасибо», «большой», «никогда», «хотеть» и др.).

В связи с критикой изданного словаря в начале 1930-х гг. М. Уэст начал разработку другой концепции лексического минимума, связанной с преподаванием английского языка как иностранного, – Reading Movement. Находясь на службе в Индии, М. Уэст интуитивно отбирал слова для обучающихся, необходимые индийцам для чтения литературы на английском языке. Вручную ему удалось отобрать около 1 800 слов. В 1934–1935 гг. он доложил итоги своей работы на международной конференции лингвистов, где выступили также другие ведущие специалисты в области преподавания английского языка: Г. Палмер, Э. Торндайк, Л. Фосетт. Ими был составлен доклад под названием *Interim Report on Vocabulary Selection* (1936), в котором эти ученые предложили избрать в качестве ведущего принципа отбора минимума лексики для освоения ее иностранцами критерий частотности. Кроме того, были выделены следующие критерии составления минимизированного словника: 1) структурная ценность (включались строевые слова языка); 2) тематическая широта (исключались термины); 3) стилистическая нейтральность (исключались разговорные слова и сленг); 4) наличие слов-определителей (включались слова, необходимые для формулирования дефиниций в словарях); 5) словообразовательная способность [Маркина, 2011].

Впоследствии на основе этих критериев в 1953 г. М. Уэст подготовил новый список [West, 1953], насчитывавший 2 000 слов с дериватами. Процесс выбора слов опирался на опыт членов комитета и включал как объективные, так и субъективные критерии. Объективный отбор был основан главным образом на списке 1 500 самых частотных слов английского языка [Fauccett, 1936], в то время как субъективные критерии были основаны на списке слов Палмера [Palmer, 1931], с одной стороны, и словаре Уэста [West, 1927], с другой, были включены слова со служебными функциями, а именно предлоги, местоимения, вспомогательные глаголы, союзы и определители. Также были включены слова, использование которых не ограничивалось определенным временем или местом (универсальность), а слова, используемые в ограниченных областях (религия, моральные понятия, собственные имена), были исключены. Слова, которые были сочтены полезными для личного и профессионального развития учащихся в возрасте от 12 до 18 лет, относились к широким предметным областям и потому вошли в список. К каждому слову были сделаны такие пометы, как отнесенность к частям речи, LCB и иллюстративный пример, а также указывалось процентное соотношение употребимости LCB в конкретном значении, что и стало отличительной особенностью *General Service List* (GSL).

Одним из базовых критериев при создании данного списка был параметр простоты лексики, отбираемой для коммуникации на начальном уровне владения языком, в результате чего отобран минимальный набор простой лексики. Однако этот критерий оказался ненадежным, поскольку слова, выделенные авторами, не всегда являлись частотными, и был необходим дополнительный объективный критерий, которым стал критерий покрываемости текстов предлагаемым списком слов. Теперь считается, что список может быть признан педагогически эффективным и надежным только при большом проценте покрытия текстов (порядка 95%). Методы, положенные в основу данного списка, до сих пор активно используются в европейской учебной лексикографии.

Дидактического направления придерживались и другие составители частотных лексических списков того времени (см. [Thorndike, Lorge 1944; Lorge 1949; Carroll et al., 1971] для английского, а также [Morgan, 1928; Bakonyi, 1934] для немецкого языка). В первый период списки составлялись на материале образовательных текстов для школьников, а с течением времени стали использоваться газетные статьи и другие печатные издания.

Основной проблемой этих и других ЛМ того времени был, прежде всего, малый объем корпусов и поверхностное применение статистики, причиной чего был, в частности, недостаток качественных вычислительных инструментов. В расчет бралась почти исключительно

абсолютная частотность слов, которая напрямую зависит от словарно-тематического состава конкретного корпуса. С чисто лингвистической точки зрения проблемой этих списков был чрезмерный крен в сторону существительных по сравнению с другими частями речи (аналогичная картина наблюдается и в более поздних статистических подходах), а также игнорирование целостности семантических полей. К примеру, в список могла попасть лишь часть наименований дней недели или месяцев. Кроме того, такие списки, как видно из приведенных выше рекомендаций, были плохо приспособлены для образовательных целей и не могли использоваться в качестве ЛМ для изучающих иностранный язык.

В то же время, как отмечается в [Маркина, 2011], все больше исследователей осознают необходимость введения содержательных критериев, которые заменили бы или, по меньшей мере, дополнили критерий частотности. Так, в 1950–1960-е гг. французскими лингвистами Ж. Гугенеймом и Р. Мишеа был предложен принцип «присутствия в сознании» (фр. *disponibilité*, англ. *availability*). Данный принцип предполагал отбор такой лексики, которая была бы полезной в повседневной жизни, поскольку она неизбежно возникает в ситуациях конкретного общения: «Мы используем слова *autobus* (автобус), *bouton* (пуговица, кнопка), *fourchette* (вилка) только тогда, когда они вызываются темой разговора» [Гугенейм, 1961]. Данный список был сформирован с помощью опроса школьников: в каждой из 16 основных тем (части тела, одежда, дом, продукты питания и т. п.) было необходимо отобрать по 20 необходимых лексических единиц. В результате обработки статистических данных был создан *Dictionnaire Fondamental de la Langue Française* [Gougenheim, 1958].

В немецкой лексикографии появляются списки, имеющие статистическую основу, но при этом также ориентированные на коммуникативные потребности пользователей. Достигалось это введением дополнительных мер, таких как включение экспертами менее частотных слов для заполнения лакун в семантических полях или устранение «лишних» частотных слов. Одним из ярких примеров применения такого подхода является список из 1 084 лемм (595 000 словоформ) Й. Пфеффера [Pfeffer, 1970]. Список был составлен автором на основе собранного им корпуса объемом 650 000 словоупотреблений, содержащего записи устной речи на повседневные темы от носителей из ФРГ, ГДР, Швейцарии и Австрии, и дополнен списком наиболее важных существительных, глаголов и прилагательных на материале опроса школьников. В списке также присутствуют собственные дополнения слов, которые, по мнению автора, должны завершить минимально необходимый лексический ряд, такие как *луна* в ряду слов *солнце* и *звезды* или слова *мотор* и *руль* в дополнение к слову *мотоцикл*.

С начала 2000-х гг. с ростом объема корпусов и формированием репрезентативных подкорпусов устной и письменной речи и частных жанрово-стилевых разновидностей языка исследователи все больше стали обращать внимание на функциональное разнообразие представленной в корпусе лексики, а развитие технологий позволило расширить инструментарий статистических метрик для выявления релевантных слов. В качестве примера можно привести список [Leech et al., 2001], составленный на основе 100-миллионного корпуса BNC. Список упорядочен по алфавиту с разбиением на лексемы и словоформы каждой лексемы и по частотности (рангам). В нем также представлена статистика использования единиц по подкорпусам (письменному и устному). Как и в работах предшественников, при составлении этого списка используется мера абсолютной частотности, измеряемая в количестве словоупотреблений на миллион слов, с нижней границей отсечения в 10 вхождений на миллион, а также разбиение на ранги. Однако с появлением возможности анализировать и сравнивать отдельные подкорпусы по регистрам, жанрам и стилям возникает необходимость соизмерять объемы подкорпусов и вырабатывать новые статистические метрики их сбалансированности. Для этого вводятся две дополнительные меры: мера характерности и мера дисперсии. Мера характерности представляет собой меру вероятности появления слова в подкорпусе, тогда как мера дисперсии определяет охват (т. е. количество секторов из 100 равных секторов в подкорпусе с данным словом) и распределение (т. е. равномерность распределения данного слова по секторам объемом в миллион слов).

Учет регистровых, жанровых и стилевых различий в сочетании с более детальной проработкой статистических метрик позволяет достичь большей точности в оценке частотности слов в языке, а также дает возможность приспособить список слов под более конкретные практические нужды. «Новая волна» статистических подходов к составлению списков наблюдается и в немецкой лингвистике. В 2005 г. Р. Джонс и Э. Чирнер собрали сбалансированный корпус объемом 4,2 млн словоупотреблений с автоматической лемматизацией и частотной разметкой [Johns, Tschirner, 2006]. На основе этого корпуса они извлекли 4 000 наиболее частотных лемм и упорядочили их по частотности и по алфавиту. В первой части списка все слова были даны вместе, с указанием части речи и с английскими переводами, примерами и важными коллокациями, релевантность которых была рассчитана по минимальной частоте на миллион слов. Во второй части списка те же слова были объединены в отдельные таблицы по тематическому принципу. Авторы списка утверждали, что их корпус покрывает от 80 до 90 % произвольного неспециализированного текста, что было позднее подтверждено посредством квантитативного анализа в работе [Winnerlöv, 2014].

В 2013 г. на основе 4 корпусов текстов: LOB (Lancaster-Oslo-Bergen Corpus), BNC (British National Corpus), BE06 (The BE06 Corpus of British English (BE06), EnTENTen12, создан список New General Service List [Brezina, Gablasova 2015] объемом в 3 000 слов. Необходимость создания нового списка общего пользования взамен по-прежнему широко используемого списка Уэста [West, 1953] обосновывается авторами, во-первых, тем, что часть лексики в его списке уже устарела (к примеру, имеется слово «телеграф», но отсутствует «компьютер»). Кроме того, используемый при его составлении критерий обязательной стилистической немаркированности отбираемой лексики игнорирует некоторые стилистически окрашенные частотные лексические единицы, необходимые даже на базовых уровнях владения языком. Итоговый объем проанализированной авторами лексики составил 12 миллиардов словоформ. В новом получившемся списке оказалось 2 494 леммы, покрывающие 80,1–81,7 % наугад выбранного неспециализированного текста.

В последнее время проблематика ЛМ активно исследуется в рамках японского проекта «Basic German Vocabulary for Foreign Language Learners: A data-driven Approach» по составлению учебных минимумов немецкого языка. В работе [Okamura et al., 2012] на основе нового объемного корпуса Basic German Corpus объемом в 845 млн словоупотреблений с двумя примерно равными подкорпусами немецкоязычных онлайн-форумов и газетных текстов производится анализ лексики с целью выделения лексического ядра с акцентом на большем объеме и большей дифференцированности статистических измерений. Вывод, к которому приходят исследователи на основе сопоставительного анализа двух подкорпусов, состоит в том, что недифференцированный подход к составлению списка лексики приводит к нежелательному смещению важных жанровых и стилевых различий текстов, что говорит о необходимости более подробного статистического анализа.

Появление корпусов большого объема и надежной статистической методологии дало возможность задаться еще одним существенным вопросом о том, насколько статистически выделимо лексическое ядро языка. Попытка поиска ответа на этот вопрос в [Scharloth et al., 2016] на материале того же Basic-German Corpus дала отрицательный результат: частотность слов монотонно убывает, и график не имеет явных изгибов, позволяющих провести деление на ядерную и остальную часть лексики. На этом основании авторы заключают, что граница между ядерной и неядерной лексикой должна проводиться исходя из конкретных практических целей составителей списка.

Таким образом, зарубежный опыт составления ЛМ показывает, что значительным методологическим шагом вперед стало активное использование текстовых корпусов, а с ростом их объема и жанрового разнообразия – учет жанрово-стилевых особенностей лексики. В немалой степени прогресс в этой области также предопределило развитие статистических методов, при помощи которых стало возможно сопоставлять между собой лексемы с точки

зрения не только частотности, но также, среди прочего, стабильности во времени и стабильности по типам текстов.

Основные различия в российских и зарубежных подходах

Как показывает настоящий обзор, преобладающей тенденцией в зарубежных подходах к составлению ЛМ является использование в качестве основы корпуса или корпусов письменных, а в последнее время также и устных текстов. В свою очередь, российские подходы в большей степени консервативны и чаще всего опираются на существующие словари и учебные пособия, хотя иногда также используют частотные списки слов. Во многом это объясняется тем, что российские ЛМ составляются с ориентацией на коммуникативные потребности иностранцев, изучающих русский язык, тогда как большинство составителей зарубежных ЛМ последнего времени ставит более общую задачу выявления лексического ядра языка, что, однако, не исключает появления конкурентных коммуникативно ориентированных и смешанных списков. Тем не менее, и в России, и за рубежом отбор слов в списки с коммуникативной ориентацией, как правило, производится по субъективным критериям, которые в недостаточной степени отражают статистические характеристики реального языкового узуса. Это видно из сравнения процента покрываемости ЛМ произвольно выбранных неспециализированных текстов на языке списка, который оказывается выше у ЛМ, составленных на статистической основе, что доказывает их большую валидность по сравнению с коммуникативно-ориентированными и смешанными ЛМ.

Результаты обзора суммированы в таблице ниже.

Сравнение российского и зарубежных подходов к созданию лексического минимума
Comparing of Russian and foreign lexical minima approaches

Показатель	Лексикография		
	русская	английская	немецкая
Предназначение	Учебные	Теоретические, общего пользования	Теоретические, общего пользования, учебные
Подход к составлению	Преимущественно коммуникативный	Преимущественно статистический	
Источники данных	Учебники, словари частотные списки ранее созданные частотные списки и ЛМ	Корпуса, словари, частотные списки, опрос носителей	
Какой процент текстов должен покрывать ЛМ	70	не менее 80	

Заключение

Приведенный в статье обзор обнаруживает целый ряд моментов, которые необходимо учитывать составителям ЛМ. Прежде всего, составление ЛМ в настоящее время практически непредставимо без корпусной основы. Корпус или корпуса, которые используются в качестве основы, должны быть достаточного объема и в достаточной мере отражать разнообразие языка с точки зрения регистров, жанров и стилей. В свою очередь, сами списки должны демонстрировать высокий процент покрытия исходного корпуса и содержать лексемы, максимально охватывающие все представленные типы текстов. К количественным характеристикам, помимо абсолютной частотности, относятся меры стабильности и продуктивности.

К качественным характеристикам относятся тематическая организация, грамматическая информация и типичные контексты употребления слова. Объем списка, как показывает практика, не играет принципиальной роли и должен определяться исходя из конкретных целей, преследуемых составителями.

Список литературы / References

- Андрюшина Н. П., Козлова Т. В.** Лексический минимум по русскому языку как иностранному. Элементарный уровень. Общее владение. М.; СПб.: ЦМО МГУ, Златоуст, 2006.
Andryushina, N. P., Kozlova, T. V. A minimal Russian wordlist for foreign learners. Elementary level. General knowledge. Moscow, St. Petersburg, Zlatoust, 2006. (in Russ.)
- Владимирова Т. Е., Нахабина М. М., Соболева Н. И., Андрюшина Н. П.** Государственный стандарт по русскому языку как иностранному. Элементарный уровень. 2-е изд., испр. и доп. М.; СПб.: Златоуст, 2001. 28 с.
Vladimirova, T. E., Nakhabina, M. M., Soboleva, N. I., Andryushina, N. P. State Standard for the Russian Language as a Foreign Language. Elementary level. 2nd ed., corr. and add. Moscow, St. Petersburg, Zlatoust, 2001, 28 p. (in Russ.)
- Гугенейм Ж.** Некоторые выводы статистики словаря // Методика преподавания иностранных языков за рубежом. М.: Прогресс, 1967. Вып. 1. С. 302.
Gughenheim, J. Some conclusions of the statistics of the dictionary. In: Methods of teaching foreign languages abroad. Moscow, Progress, 1967, vol. 1, p. 302. (in Russ.)
- Дунаева Л. А., Соррилъя К. М.** Стандартный лексический минимум в электронном практикуме по лексике русского языка // Русский язык за рубежом. 2010. № 6. С. 22–28.
Dunaeva, L. A., Sorrilja, K. M. A standard minimal Russian wordlist for foreign learners in the online platform for learning Russian vocabulary. *Russian Language Abroad*, 2010, no. 6, p. 22–28. (in Russ.)
- Засорина Л. Н.** Частотный словарь русского языка: Около 40 000 слов / Под ред. Л. Н. Засориной. М.: Рус. яз., 1977. 936 с.
Zasorina, L. N. Frequency dictionary of the Russian language: About 40 000 words. Ed. by L. N. Zasorina. Moscow, Rus. lang., 1977, 936 p. (in Russ.)
- Костомаров В. Г.** Принципы отбора лексического минимума // Русский язык в национальной школе. 1963. № 1. С. 31.
Kostomarov, V. G. Principles of creating a minimal wordlist. *Russian Language in the National School*, 1963, no. 1, p. 31. (in Russ.)
- Лённгрен Л.** Частотный словарь современного русского языка. Uppsala, 1993.
Lyonngrén, L. Frequency Dictionary of Modern Russian Language. Uppsala, 1993.
- Маркина Е. И.** Лингводидактические основы разработки лексических минимумов по русскому языку как иностранному (для разных уровней и профилей обучения): Автореф. дис. ... канд. пед. наук. М., 2011. 24 с.
Markina, E. I. Linguodidactic basis for creating minimal wordlists for Russian foreign learners. Abstract. Moscow, 2011. 24 p. (in Russ.)
- Маркина Е. И., Соррилъя К. М.** Основные подходы к минимизации лексики в российской и европейской учебной лексикографии // Вестник РУДН. Серия: Вопросы образования: языки и специальность. 2011. № 3. С. 77–84.
Markina, E. I., Sorrilja, K. M. Main approaches to vocabulary minimization in the Russian and the European lexicography. *Vestnik RUDN. Series: Issues in education: languages and profession*, 2011, no. 3, p. 77–84. (in Russ.)
- Морковкин В. В., Дорогонова И. А.** О новом типе лексического минимума современного русского языка // Русский язык за рубежом. 1976. № 2. С. 59–62.
Morkovkin, V. V., Dorogonova, I. A. On the new type of a minimal modern Russian wordlist. *Russian Language Abroad*, 1976, no. 2, p. 59–62. (in Russ.)
- Трушина Л. Б.** Русский язык как иностранный в специальных целях: прошлое, настоящее, будущее. Опыт исторического обзора / Под ред. В. Г. Костомарова. М., 2013. С. 8–9.
Trushina, L. B. Russian as a foreign language for special purposes: past, present and future. A historical overview. Ed. by V. G. Kostomarov. M., 2013. P. 8–9. (in Russ.)

- Штейнфельдт Э. А.** Частотный словарь современного русского литературного языка. Таллин, 1963.
- Steinfeld, E. A.** A frequency dictionary of modern Russian literary language. Tallinn, 1963. (in Russ.)
- Bakonyi, H.** Die gebräuchlichsten Wörter der deutschen Sprache für den Fremdsprachenunterricht stufenmäßig zusammengestellt. München, 1934.
- Brezina, V., Gablasova, D.** Is there a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, 2015, vol. 36/1, p. 1–22.
- Carroll, J. B., Davies, P., Richman, B.** The American Heritage word frequency book. Boston, Houghton Mifflin, 1971.
- Ek, J. A. van.** Scope. Levels. Strasbourg, Council of Europe, 1987, vol. 2.
- Faucet, L., West, M., Palmer, H., Thorndike, E. L.** The Interim Report on Selection of the English Language as a Foreign Language. London, 1936.
- Gougenheim, G.** Dictionnaire fondamental de la langue française. Paris, Didier, 1958.
- Johns, R., Tschirner, E.** A frequency dictionary of German: Core vocabulary for learners. Routledge, 2006.
- Leech, G., Rayson, P., Wilson, A.** Word Frequencies in Written and Spoken English: Based on the British National Corpus. London, Longman, 2001.
- Lorge, I.** Semantic count of the 570 commonest English Words. New York, Columbia University Press, 1949.
- Morgan, B. Q.** German Frequency Word Book. New York, Macmillan, 1928.
- Ogden, C. K.** Basic English, London. Kegan Paul, Trench Trubner & Co., Ltd. 1930.
- Okamura, S., Lange, W., Scharloth, J.** Methoden der Bestimmung des Kernwortschatzes Deutsch. Grundwortschatz Deutsch: Lexikografische und fremdsprachendidaktische Perspektiven. Tokyo, Studienreihe der Japanischen Gesellschaft für Germanistik, 2012, P. 29–44.
- Palmer, H.** Second interim report on vocabulary selection submitted to the Eighth Annual Conference of English Teachers under the auspices of the Institute for Research in English Teaching. Tokyo, IRET, 1931.
- Pfeffer, J. A.** Grunddeutsch. Basic (Spoken) German Dictionary. Englewood Cliffs, 1970.
- Scharloth, J., Okamura, S., Lange, W.** Gibt es einen Kernwortschatz? Datengeleitete Perspektiven auf die Erstellung von Grundwortschatzen für Deutsch als Fremdsprache. Brunetti, S. u.a. (Hrsg.). Versprachlichung von Welt – Il mondo in parole. Festschrift zum 60. Geburtstag von Maria Lieber. Stauffenburg, 2016.
- Thorndike, D. L., Lorge, I.** The Teacher's Word Book of 30,000 Words. New York, NY, Bureau of Publications, Teachers College, Columbia University, 1944.
- West, M.** A general service list of English words. London, Longman, 1953.
- West, M.** The new method readers (new series). Bombay and Calcutta, Longmans, Green, 1927.
- Winnerlöv, J.** Wie viele Wörter muss man kennen, um einen deutschen Text zu verstehen? Eine quantitative und qualitative Untersuchung zur lexikalischen Textdeckung. Stockholms Universitet, 2012.

*Материал поступил в редколлегию
Date of submission
24.10.2018*

Сведения об авторах / Information about the Authors

Муравьев Никита Алексеевич, преподаватель Школы лингвистики Национального исследовательского университета «Высшая школа экономики» (ул. Мясницкая, 20, Москва, 101000, Россия)

Nikita A. Muravyev, Lecturer of the Faculty of Humanities, School of Linguistics, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow, 101000, Russian Federation)

nmuravev@hse.ru

ORCID 0000-0001-9586-5928

SPIN 8474-3643

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