

Введение к новому частотному словарю русской лексики

С.А. Шаров, О.Н. Ляшевская

1 Введение

Частотный словарь является источником информации о том, какие слова более употребительны в языке, а какие менее частотны. Он содержит списки слов, при которых указывается, с какой частотой они встречаются в текстах. Для того, чтобы этот показатель был более достоверным, частотность слова подсчитывается на основе большого корпуса текстов.

Авторы словаря английского языка «Word Frequencies in Written and Spoken English» (Leech et al. 2001) сравнивают частотный словарь с телефонной книгой. Действительно, такие словари состоят главным образом из заглавных слов и списков чисел. Однако частотный словарь находит своего «читателя», поскольку собранная в нем информация необходима для многих целей: обучение языку, лингвистические исследования, выбор словников для создания словарей, компьютерные приложения, например, информационный поиск или системы фильтрации спама.

Для русского языка было разработано несколько частотных словарей: Э.А. Штейнфельд (1963), Л.Н. Засориной (1977), Л. Леннгрена (1993) и др., но все эти словари были созданы на основе относительно небольших коллекций текстов (400 тысяч - 1 миллион слов) и в большой степени отражают специфику русского языка советского периода: частоты слов *товарищ* и *партия* в них сопоставимы со служебными словами.

Отдельную отрасль статистических словарей составляют словари языка Пушкина, Достоевского, Грибоедова, Цветаевой (Виноградов 1956-1961, Шайкевич и др. 2003, Поляков 1999, Белякова и др. 1996), которые полностью описывают язык данного писателя. В нашем частотном словаре представлен срез всего потенциально бесконечного множества текстов, функционирующих в современном русском языке.

Некоторые частотные словари, например, Davies 2005, были созданы для преподавания языка. В них отражены не только частотность отдельных слов, но и приводится дополнительная грамматическая и лексическая информация. Такая работа для русского еще предстоит. Целью создания данного словаря было предложить пользователям достаточно представительный базовый словник современного русского языка, который можно использовать и адаптировать для разнообразных целей.

2 О Национальном корпусе русского языка

Данный словарь основан на выборке текстов Национального корпуса русского языка, представляющей современный русский язык периода 1950-2007 годов. Объем выборки составляет около 100 млн. словоупотреблений.

Национальный корпус русского языка (www.ruscorgpora.ru) является результатом большого проекта РАН (Шаров 2003, Плунгян 2005 и др.), начатого в 2001 году. В настоящее время Национальный корпус включает электронные тексты письменного русского языка XVIII — начала XXI века, записи устной речи, поэтический и диалектный корпусы, корпус древнерусского языка, а также параллельный русско-английский и русско-немецкий корпус.

Для того, чтобы корпус мог предоставить достоверные данные о частотности слов в языке, он должен быть большим по объему и представительным по охвату материала, т. е. содержать тексты разных жанров и стилей в определенной пропорции. В этом отношении Национальный корпус русского языка соответствует лучшим образцам национальных корпусов, созданных для разных языков мира, таким как Британский национальный корпус (British National Corpus), Корпус испанского языка (Corpus del español), Чешский национальный корпус (Český národní korpus). Тексты в корпусе, а также отдельные элементы текста (словоформы, знаки препинания,

абзацы и т. п.) особым образом аннотированы. Для задач составления частотного словаря особой ценностью обладают метатекстовая и лексико-грамматическая разметка.

Первый вид аннотации содержит информацию об авторе текста, его поле и возрасте, о названии текста и времени его создания, а также о типе и жанровой принадлежности. В НКРЯ тексты классифицируются по нескольким параметрам (Савчук 2005). Художественной литературе приписываются атрибуты жанра (фантастика, историческая проза), типа (роман, рассказ), времени и места описываемых событий. Нехудожественные тексты делятся на 9 групп по сфере применения: бытовая, официально-деловая, производственная, публицистика, устные тексты, реклама, учебно-научная, художественная и церковно-богословская. В дополнение к этому используется открытый список типов текстов, например, интервью, отзыв, закон (в настоящее время список содержит более 100 типов). Тематика текстов кодируется списком из 52 категорий, имеющих разную степень подробности: от «экономика» или «внутренняя политика» до «безопасность», «путешествия» или «вооруженные конфликты».

Табл. 1. Функциональные стили подкорпуса современного русского языка

Художественная литература	36,38%
Нехудожественная литература, в т. ч.	41,65%
публицистическая	17,03%
учебно-научная	11,38%
официально-деловая	1,62%
церковно-богословская	1,42%
электронная коммуникация	1,36%
реклама	0,55%
бытовая	0,46%
производственно-техническая	0,25%
Устная литература, в т. ч.	4,93%
устная публичная	4,02%
устная непубличная	0,35%
кино	0,57%

Метатекстовая разметка дает возможность поддерживать в корпусе выверенный баланс текстов разных типов, см. табл. 1. На основе метатекстовой информации можно строить частотные списки на отдельных выборках корпуса и сравнивать их между собой.

Другой вид разметки, лексико-грамматическая, позволяет установить исходную форму слова (лемму), ее часть речи и такие грамматические характеристики, как падеж, число, время и т. д. Это дает возможность собрать данные о частотности не только отдельных словоформ, но и лексем, а также об употребительности тех или иных грамматических категорий. При создании настоящего частотного словаря был использован вариант лексико-грамматической разметки корпуса с автоматическим разрешением морфологической омонимии, см. ниже.

3 Размер корпуса и надежность выборки

Существующие частотные словари для русского языка были построены на сравнительно небольших корпусах (400 тыс. словоупотреблений для словаря Штейнфельд, один миллион для словарей Засориной и Леннгрена): ЭВМ первых поколений не могли работать с корпусами большего размера. Интересно, что теоретические рекомендации, выработанные в 1970-е годы (Пиотровский и др. 1972), также доказывали, что для достоверного описания 1600-1700 наиболее частотных слов достаточно использовать корпус размером 400 тыс. словоупотреблений. Эта аргументация строилась на понятии доверительного интервала, который широко используется в

статистике и социологии: если мы знаем размер выборки и экспериментальную вероятность события в этой выборке (т.е. частоту слова в нашем корпусе), то мы можем вычислить доверительный интервал вероятности этого события на всей популяции (т.е. частоту употребления того же слова во всем пространстве языка).

В Таблице 2 приводятся примеры частоты отдельных слов в словарях Леннгрена, Засориной и Штейнфельд в сравнении с частотами НКРЯ и 150-миллионного корпуса русского языка, собранного из Интернета (о последнем см. Sharoff 2006). Несмотря на то, что слова *думать*, *задача*, *любить* безусловно относятся к ядру языка (входят в число 200-500 самых частотных лемм), в небольших корпусах даже их частота различается весьма существенно. Частота сравнительно менее частотных слов (*загрязнение*, *изучение*, *милый*) варьируется в еще больших пределах. Хотя состав Интернет-корпуса довольно существенно отличается от НКРЯ (большим количеством технических текстов и форумов и меньшим количеством художественной литературы), различия в частоте этих единиц между ними не столь велики.

Табл. 2: Сравнение частоты отдельных слов (среднее на миллион словоупотреблений)

Лемма	Леннгрен	Засорина	Штейнфельд	НКРЯ	Интернет
Власть	202	364	138	422	428
Думать	609	1094	1058	865	818
загрязнение	69	1	0	9	11
задача	499	421	250	228	292
изучение	193	110	0	63	78
любить	415	632	595	549	650
милый	58	242	135	129	ПО

Как видим, теоретические рекомендации относительно достаточного размера корпуса в данном случае оказываются не слишком достоверными. Причина этого кроется в исходных допущениях на нормальное Гауссово распределение частоты слов, в соответствии с которым каждое слово встречается с одинаковой частотой во всех текстах. Если слово встретилось в тексте один раз, то при нормальном распределении это не влияет на вероятность его употребления там во второй раз. Но в реальности это не так. Каждый текст имеет некоторую собственную тему, слова которой в этом тексте будут употребляться намного чаще среднего. В тексте про хоббитов слово *хоббит* будет употребляться так же часто, как и многие служебные слова, что существенно повысит его частоту в корпусе, который будет включать хотя бы один такой текст.¹ В результате частотный список, построенный на основе корпуса, отражает специфику тех текстов, которые попали в него при его составлении.

Таблица 2 показывает несовершенство частотных словарей, построенных на относительно небольших корпусах, но простое увеличение размера корпуса также не гарантирует стабильности результатов. При интерпретации списков частотного словаря надо помнить, что любой корпус, каким бы большим он ни был, является конечным подмножеством потенциально бесконечного множества текстов на данном языке. Любая другая выборка этого подмножества породит несколько другой список, который будет отличаться в своих менее частотных элементах. Корпус большего размера, отражающий большее количество тем и функциональных стилей (корпус типа BNC или НКРЯ), обеспечивает хорошую надежность для наиболее частотных элементов. Тем не менее, дальнейшее увеличение объема текстов в ущерб их разнообразию (см., например, проекты создания Гига-корпусов английского и китайского языков, содержащих более миллиарда словоупотреблений новостных текстов, Cieri & Liberman 2002), может приводить к меньшей надежности частотного списка на таких корпусах за счет сдвига их словаря в сторону новостной лексики.

Поскольку задачей частотного словаря является не просто ранжировать слова по их частоте в отдельном корпусе, но и определить лексическое ядро языка, необходимо отделить слова, часто встречающиеся во многих текстах, от тех, чье лексическое поведение подобно словам *Норвега* или *хоббит*, и которые случайно оказались в той или иной позиции частотного списка. Так в Чешском национальном корпусе используется понятие средней уменьшенной частоты (ARF,

¹ Кеннет Черч называл эту ситуацию проблемой Норвеги (Church 2000), Адам Килгаррифф - whelk problem, от сравнительно редкого английского слова, обозначающего вид моллюска (Kilgarriff 1997).

Average Reduced Frequency), в котором частота слова взвешивается по расстоянию между отдельными словоупотреблениями (Čermak & Křen 2005). Во многих частотных словарях (Леннгрена, Британского национального корпуса, словаря французской лексики в области бизнеса) используется коэффициент D , введенный А. Жуйаном (Juilland et al. 1970), который принимает во внимание как число документов, в которых встречается слово, так и его относительную частоту в этих документах:

$$D = 100 \times \left(1 - \frac{\sigma}{\mu\sqrt{n-1}}\right)$$

где μ – средняя частота слова по всему корпусу, σ – среднее квадратичное отклонение этой частоты на отдельных документах, n – число документов, в которых встречается это слово.

Значение D у слов, встречающихся в большинстве документов, близко к 100, а у слов, часто встречающихся лишь в небольшом числе документов, близко к 0.² Частотный список словаря Леннгрена даже отсортирован по значению произведения этого коэффициента на среднюю частоту слова. В связи с тем, что теоретический статус этого произведения неясен, мы не считали целесообразным сортировать наш словарь по нему. Однако его указание для каждого слова дает возможность оценить, насколько оно специфично для отдельных предметных областей. Например, слова *жуткий*, *специфический* и *сырье* имеют примерно равную частоту (21 ipm), но при этом коэффициент D у *специфический* - 0.66, *сырье* - 0.18, а у *жуткий* - 0.78, что означает, что последнее слово значимо для большего числа предметных областей и (при прочих равных условиях) имеет большие шансы на место в неспециализированном словаре.

4 Принципы создания словника

4.1 Размер словаря

Хорошо известно, что распределение частот лексических единиц крайне неравномерно: очень небольшое количество слов встречается достаточно часто, а частота большинства обыденных слов очень невелика. Закон Ципфа (Zipf 1935) определяет обратно-пропорциональную зависимость между порядковым номером слова в частотном списке (r , ранг) и его частотой (f):

$$f \approx kr^{-\alpha},$$

где k – константа, зависящая от корпуса (абсолютное число употреблений самого частотного слова), а α – близкий к единице степенной параметр, зависящий от грамматического строя языка (следует отметить, что это эмпирическая зависимость, а не строгое математическое соответствие; более точное описание модели см. в Арапов и др. 1975). Схематически эту зависимость можно изобразить графиком на рисунке 1: частота подавляющего большинства слов очень невелика и частота более редких слов медленно падает с увеличением размера словника.

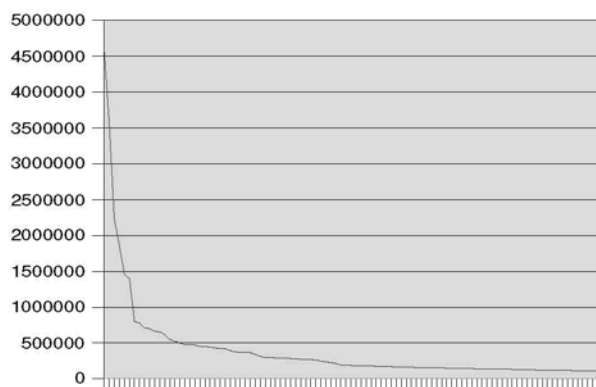


Рис. 1: Зависимость частоты от ранга (закон Ципфа)

В результате списки частотных словарей приходится ограничивать некоторым порогом, зависящим от полиграфических возможностей издания, а также вариативностью частоты в

² Здесь мы следуем методу презентации в Leech et al. 2001. В оригинале коэффициент считается без умножения на 100.

зависимости от размера корпуса. Практика создания частотных словарей на материале 100-миллионных корпусов обычно ограничивает частотные списки словами с частотой около 5 употреблений на миллион слов, что в 135 миллионах слов НКРЯ дает около 675 употреблений самых редкочастотных слов в списке (и дает словник около 14,000 слов). За пределами этого списка частота слова становится менее предсказуемой. Так, по материалам НКРЯ слова *неумолимо* и *подвох* входят в число 20000 самых частотных слов, а слова *изворотливый* и *раскуривать* находятся за пределами списка первых 40000, что не вполне соответствует интуиции.

Необходимо также отметить, что ответ на вопрос о размере корпуса не всегда однозначен. Под количеством словоупотреблений понимается количество элементов, полученных в результате так называемой токенизации, разбиения потока текста на элементы (токены), которые включают орфографические слова, числа и знаки пунктуации. В соответствии с разными подходами под размером корпуса можно понимать общее количество токенов, количество токенов за исключением пунктуации или количество орфографических слов. В последнем случае *двадцать пять* считается двумя словами, а *25* – одним. Иногда учитываются только слова, записанные кириллицей. При автоматическом подсчете орфографических слов также остается неопределенность в том, как учитывать разбиение дефисами и знаками переноса (ср. *как-нибудь*, *еврей-крестьянин*, 1970-е), косой чертой (*и/или*, *км/ч*), как выделять и учитывать обороты (*в течение*, *незвизая на*) и т.п. В соответствии с моделью, использованной при создании данного словаря, он основан на корпусе из 153 миллиона токенов, 135 миллионов орфографических слов, включая числа, записанные римскими и арабскими цифрами.

С точки зрения словарного запаса в НКРЯ содержится 2 322 092 отдельных орфографических словоформ, 739 930 лемм, из которых 360 755 встречаются более одного раза (268 106 встречаются более двух раз, 125 688 более десяти раз). Десять самых частотных лемм покрывают 17% текста, 100 лемм – 38%, 2000 лемм – 70% всех текстов.

4.2 Лемматизация и частеречная аннотация

Основной единицей частотного списка является лемма, или исходная форма слова. Каждой лемме приписывается информация о части речи. Выделяются следующие классы:

S — существительное (*яблоны, лошадь, корпус, вечность*)

A — прилагательное (*коричневый, таинственный, морской*)

NUM — числительное (*четыре, десять, много*)

ANUM — числительное-прилагательное (*один, седьмой, восьмидесятый*)

V — глагол (*пользоваться, обрабатывать*)

ADV — наречие (*сгоряча, очень*)

PRAEDIC — предикатив (*жаль, хорошо, пора*)

PARENTH — вводное слово (*кстати, по-моему*)

SPRO — местоимение-существительное (*она, что*)

APRO — местоимение-прилагательное (*который, твой*)

ADVPRO — местоименное наречие (*где, вот*)

PRAEDICPRO — местоимение-предикатив (*некого, нечего*)

PR — предлог (*под, напротив*)

CONJ — союз (*и, чтобы*)

PART — частица (*бы, же, пусть*)

INTJ — междометие (*уввы, батюшки*)

Еще один класс составляют части слов (COM), графически отделенные в тексте от остального слова, например, *водо* в «*водо- и теплоснабжение*», *пере* в «*пере- и недоопределение*»; *го* в «*1-го*», *летний* в «*15-летний*».

Морфологический стандарт НКРЯ, предписывающий правила лексико-грамматической аннотации словоформ (Ляшевская и др. 2005), в общем и целом руководствуется принципами Грамматического словаря русского языка (Зализняк 1977). Некоторые особенности лемматизации связаны с тем, что сбор данных происходит по преимуществу в автоматическом режиме.

Учитывается только пословная разметка: устойчивые обороты, составные предлоги и другие неоднословные лексические единицы (ср. *Новый год, в течение, тем не менее, друг друга*) не включаются в словарь. Части сложных слов, записанные через дефис (ср. *город-спутник, член-*

корреспондент), как правило, учитываются по отдельности. Исключение составляет закрытое множество лемм из словаря Зализняк 1977 с неизменяемой первой частью: *вице-президент, штаб-квартира, кисло-сладкий, из-за, еле-еле, по-турецки, кое-какой* и т. д., а также местоимения с формантами *-то, -либо, -нибудь*.

Формы причастий входят в парадигму глагола. Возвратные и невозвратные глаголы, глаголы совершенного и несовершенного вида считаются отдельными единицами словаря.

Русский язык с присущим ему богатым словоизменением создает дополнительные трудности для составителей частотного словаря, так как многие словоформы омонимичны (ср. словоформу *стали* как форму глагола *стать* и существительного *сталь*, словоформу *банка*, представляющую леммы *банк* и *банка*, слова *вера* и *Вера*). Тем не менее, для работы со статистикой исходная форма должна быть приписана любой словоформе однозначно.

В словарях предшествующего поколения (Засорина 1977, Леннгрен 1993) омонимия разрешалась вручную, так как объем обрабатываемого корпуса был незначителен. Очевидно, что для 100-миллионного корпуса такое решение не подходит. При составлении настоящего словаря был учтен опыт чешских коллег, которым пришлось дорабатывать морфологический анализатор, пополнять словарь и проводить ручную редактуру. Первоначально корпус НКРЯ был размечен морфологическим анализатором Mystem (Сегалович & Маслов 1998). Неоднозначность в лексико-грамматической разметке была разрешена с помощью программы А.В. Сокирко, использующей модель триграмм и тренировочный подкорпус со снятой вручную омонимией (Сокирко & Толдова 2005). Поскольку автоматическое разрешение омонимии допускает определенную, хотя и незначительную, погрешность, омонимы, входящие в первые 20 тысяч частотных слов, подверглись дополнительной ручной проверке.

Лексические омонимы типа *лук¹ – лук², повезти¹ – повезти²*, т. е. слова одной части речи, с одним типом словоизменения, но имеющие разные значения, в словаре не различаются. В частности, считаются одной единицей слова, различающиеся местом ударения, а также буквами *е* и *ё* (ср. *за́мок – замо́к, надеж – надёж*). Вместе с тем, слова в парах *вера – Вера, прус – Прус, су-СУ* и др. считаются двумя разными леммами: это обусловлено тем, что имена собственные и сокращения выделяются в словаре в особый список.

Омонимичные леммы, принадлежащие разным частям речи, приводятся отдельно. Исключение делается для слов неизменяемых частей речи:

печь	S	14.70	4.83
печь	V	6.27	74.86
вроде	PR,PART	113.15	3.14

Отдельная проблема для лемматизации – словоформы, которые не входят в грамматический словарь программы автоматического анализа текста, например, новые слова (*неприватизированный*), имена собственные (*Байкал*), нестандартные формы склонения и спряжения (*ходят*). При разметке корпуса анализатором Mystem доля несловарных словоформ составила 3% всех словоупотреблений и 45% списка словоформ. Леммы несловарных слов были определены с помощью программ пост-обработки морфологической разметки НКРЯ, составленных Б.П. Кобрицовым и Г.К. Бронниковым (см. подробнее Ляшевская 2007), а затем выверены вручную.

Нестандартные варианты словоизменения и искаженные написания учитываются при подсчете употреблений леммы наряду со стандартными формами склонения и спряжения.

Сокращения, которые пишутся со строчной буквы и с точкой на конце, расшифровываются: например, леммами слов *рис., тел., стр.* считаются, соответственно, *рисунок, телефон, страница* и *строение* (форма *стр.* требует разрешения омонимии).

В словаре данные о сокращенных вариантах написания приводятся курсивом после леммы, Цифра показывает долю сокращений относительно абсолютной частоты леммы во всем корпусе современного русского языка:

смотреть	V	80.61	4.91
см.	4%		

5 Структура словаря

Словарь состоит из следующих разделов:

I. Общая лексика

- алфавитный список лемм
- частотный список лемм
- распределение лемм по функциональным стилям:
 - частотный словарь художественной литературы, словарь значимой лексики художественной литературы
 - частотный словарь публицистики, словарь значимой газетно-новостной лексики
 - частотный словарь другой нехудожественной литературы, словарь значимой лексики
 - частотный словарь устной речи, словарь значимой лексики устной речи
- алфавитный список словоформ

II. Общая лексика: части речи

- частотный список имен существительных
- частотный список глаголов
- частотный список имен прилагательных
- частотный список наречий и предикативов
- частотный список местоимений (местоимения-существительные, прилагательные, наречия, предикативы)
- частотный список лемм служебных частей речи

III. Вспомогательные таблицы

- данные о частотности частеречных классов и другая статистическая информация

IV. Имена собственные и аббревиатуры

- алфавитный список лемм

В алфавитном списке лемм приводится имя леммы, часть речи, общая частота леммы, число документов, в которых она встретилась и коэффициент вариации D . Общая частота характеризует число употреблений на миллион слов корпуса, или ipm (instances per million words). Это делается для того, чтобы упростить сравнение частоты слова в разных корпусах, которые могут довольно сильно отличаться по своим размерам. Например, если слово *власть* встречается 55 раз в корпусе размером 400 тыс. слов, 364 раза в миллионном корпусе и 40598 раз в 100-миллионном корпусе современного русского языка и 55673 раза в большом 135-миллионном корпусе НКРЯ, то его частота в ipm составит 137.5, 364.0, 372.06 и 412.39, соответственно. Алфавитный список включает 50 000 наиболее частотных лемм.

В списке лемм, упорядоченном по частотности, указываются имя леммы, часть речи, общая частота леммы, число документов, коэффициент D и распределение частотности по десятилетиям. Частотный список включает 20 000 самых частотных лемм.

Частотные словари функциональных стилей составлены на основе подкорпусов художественной литературы, публицистики, другой нехудожественной литературы и устной речи. В список включены 5 000 самых частотных лемм этих подкорпусов. Список наиболее типичных лемм для каждого типа текстов был выделен на основе сравнения частоты лемм в таких текстах и в остальном корпусе. В качестве метрики сравнения был использован критерий отношения правдоподобия (log-likelihood), вычисляемый на основе следующей матрицы:

	Подкорпус	Другие тексты	Весь корпус
Частота	a	b	a+b
Размер	c	d	c+d

На основе этой матрицы значение отношения правдоподобия $G2$ можно вычислить как (Rayson & Garside 2000):

$$G2 = 2(a \ln(\frac{a}{E1}) + b \ln(\frac{b}{E2})); \text{ где } E1 = c \frac{a+b}{c+d}; E2 = d \frac{a+b}{c+d}$$

Словари значимой лексики включают по 500 лемм.

Алфавитный список словоформ включает все словоформы корпуса с частотой выше 1 ipm (всего около 15 тыс.) и содержит информацию об их общей частоте. Омонимичные словоформы помечаются знаком *.

В разделе «Части речи» частотный список лемм разбит на шесть подсписков: имена существительные, глаголы, имена прилагательные, наречия и предикативы, местоимения и служебные части речи. Здесь для каждой леммы указана ее общая частота и ранг (порядковый номер) в общем списке. Каждый список содержит по 1 тысяче наиболее частотных лемм.

Вспомогательные таблицы включают в себя данные о частотности частеречных классов, других грамматических категорий, а также информацию о покрытии текста лексемами, средней длине слова, словоформы и предложения.

Завершает словарь алфавитный список имен собственных и аббревиатур. Имена собственные отделены от основной части словника, так как образуют значительно менее стабильную в статистическом отношении группу, а их частотность в большой степени зависит от выбора текстов в корпусе и их хронотопа. В Леннгрен 1993 высказано мнение, что включение имен собственных в частотный словарь на общих основаниях неизбежно приводит к его преждевременному устареванию.

Для получения списка имен собственных и аббревиатур из конкорданса корпуса были выделены имена существительные и сокращения, написание которых в текстах с большой буквы превышало 95-процентный порог, ср. *Россия, Смирнов, ГРЭС, МИД, КЗоТ*.³ В словарь включена ядерная часть этого списка, насчитывающая 3 000 наиболее частотных единиц.

Список литературы

Арапов М.В., Е.Н. Ефимова, Ю.А. Шрейдер (1975). *О смысле ранговых распределений*. Научная и техническая информация. Серия 2. № 1. С. 9-20. <http://kudrinbi.ru/public/442/>.

Белякова И.Ю., И.П. Оловянникова, О.Г. Ревзина (сост.) (1996). *Словарь поэтического языка Марины Цветаевой*. В 4-х томах. М: Дом-музей Марины Цветаевой.

Виноградов В.В. (отв. ред.) (1956-1961). *Словарь языка Пушкина*. Т. I – IV. М.

Зализняк А.А. (1977). *Грамматический словарь русского языка: Словоизменение*. М.; 4-е изд.: М.: Русские словари, 2003.

Засорина Л.Н. (ред.) (1977). *Частотный словарь русского языка*. Москва: Русский язык.

Леннгрен Леннарт (ред.) (1993). *Частотный словарь современного русского языка* (Lönngren, Lennart. The Frequency Dictionary of Modern Russian). Acta Univ. Ups., Studia Slavica Upsaliensia Uppsala 32. Uppsala.

Ляшевская О.Н. (2007). Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика 2007. М. (в печати).

НКРЯ: *Национальный корпус русского языка 2003-2005: Результаты и перспективы*. М.: Индрик, 2005.

Пиотровский Р.Г., К.Б. Бектаев, А.А. Пиотровская (1972). *Математическая лингвистика*. М.: Высшая школа.

Плунгян В.А. (2005). Зачем нужен Национальный корпус русского языка? // *Национальный*

³ Специально отметим, что прилагательные типа *Христов, Петин, Костромской/костромской* относятся к общей лексике.

Корпус Русского Языка 2003-2005. Результаты и перспективы. М.: Индрик. С. 6-20.

Поляков А.Е. (1999). Электронный словарь языка писателя (на примере языка А.С. Грибоедова) // *Труды Международного семинара Диалог-99 по компьютерной лингвистике и ее приложениям. Таруса, 1999.* М. Т. 2. С. 230-236.

Савчук С.О. (2005). Метатекстовая разметка в Национальном корпусе русского языка // *Национальный Корпус Русского Языка 2003-2005. Результаты и перспективы.* М.: Индрик. С. 62-88.

Сегалович И., Маслов М. (1998). Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // *Труды международной семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Казань, 1998.* Т.2. С. 547–552.

Сокирко А.В., Толдова С.Ю. (2005). Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // *Интернет-математика-2005.* М.

Степанова Е.М. (1976). *Частотный словарь общенаучной лексики.* М.

Шайкевич А.Я., Андрищенко В.М., Ребецкая Н.А. (2003). *Статистический словарь языка Достоевского.* М.: Языки славянской культуры.

Шаров С.А. (2003). *Представительный корпус русского языка в контексте мирового опыта* // Научная и техническая информация. Серия 2. № 5. С. 8-19.

Штейнфельд Э.А. (1963). *Частотный словарь современного русского литературного языка.* Таллин.

Čermák, František & Michal Křen (2005). New generation corpus-based frequency dictionaries: The case of Czech // *International Journal of Corpus Linguistics*, 10. P. 453-467.

Church, Kenneth W. (2000). Empirical estimates of adaptation: the chance of two Noriegas is closer to $p/2$ than p^2 // *Proceedings of the 18th Conference on Computational Linguistics (COLING), Saarbrücken, Germany, 2000.* Vol. 1. P. 180-186.

Cieri, Christopher & Mark Liberman (2002). Language resources creation and distribution at the Linguistic Data Consortium // *Proceedings of LREC02.* Las Palmas, Spain, 2002. С. 1327-1333.

Davies, Mark (2005). *A Frequency Dictionary of Spanish: Core Vocabulary for Learners.* London – N.Y.: Routledge.

Josselson, Harry H. *The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian.* Detroit: Wayne University Press, 1953.

Juilland, Alphonse, Dorothy Brodin & Catherine Davidovitch (1970). *Frequency Dictionary of French Words.* The Hague—Paris: Mouton.

Kilgarriff, Adam (1997). Putting frequencies in the dictionary // *International Journal of Lexicography*, 10 (2). P. 135-155.

Leech, Geoffrey, Paul Rayson & Andrew Wilson (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* Longman, London.

Rayson, Paul & Roger Garside (2000). Comparing corpora using frequency profiling // *Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000.* P. 1-6.

Sharoff, Serge (2006). Creating general-purpose corpora using automated search engine queries // Baroni, Marco, Silvia Bernardini (eds.): *WaCky! Working papers on the Web as Corpus.* Bologna: Gedit. <http://wackybook.sslmit.unibo.it>.

Zipf, George Kingsley (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Boston: Houghton Mifflin.