Введение к частотному словарю современного русского языка

С.А. Шаров, О.Н. Ляшевская

1 Введение

Частотный словарь служит 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, частота слова подсчитывается на основе большого корпуса текстов.

Авторы частотного словаря английского языка «Word Frequencies in Written and Spoken English» (Leech et al. 2001) начинают свое введение сравнением его 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 решения многих задач в самых разных областях. Это, например, обучение языку, лингвистические научные исследования, составление словарей, а также компьютерные приложения, в частности, информационный поиск и 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 Грибоедова, Достоевского, Цветаевой (Поляков 1999, Шайкевич и др. 2003, Белякова и др. 1996), которые полностью описывают язык одного писателя. Существуют также специализированные словари, отражающие язык художественной литературы (Josselson 1953), науки (Степанова 1970), публицистики (Шайкевич и др. 2008). Корпус, на котором основан наш частотный словарь, включает тексты самых разных авторов; помимо литературных произведений, в него входит публицистика и другие жанры нехудожественной литературы, а также записи устной речи – то есть в словаре представлен срез всего потенциально бесконечного множества текстов, функционирующих в современном русском языке.

Некоторые частотные словари создаются специально для преподавания языка. например, испанский словарь Davies 2005 или словарь американского английского Davies & Gardner 2010. В них отражены не только частота отдельных слов, но и приводится дополнительная грамматическая и лексическая информация, а также типичные коллокации (словосочетания). Для русского языка такая работа 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

Словарь основан на коллекции текстов Национального корпуса русского языка, представляющей современный русский язык периода 1950–2007 годов. Объем выборки, на которой строится большинство разделов словаря, составляет 92 млн. словоупотреблений.

Национальный корпус русского языка (http://www.ruscorpora.ru) является результатом большого проекта РАН (Шаров 2003, Плунгян 2005, НКРЯ 2006–2008 и др.), начатого в 2001 году. В настоящее время Национальный корпус включает электронные тексты письменного русского языка XVIII – начала XXI века (основной корпус), записи устной речи, газетный, поэтический и диалектный корпусы, синтаксически размеченный корпус, акцентологический, мультимедийный и обучающий корпусы, корпус древнерусского языка, а также параллельный русско–английский и русско–немецкий корпус.

Для того чтобы корпус мог предоставить достоверные данные о частоте 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) и др. 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 (имя, пол и год рождения), 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 и т.п.), типа (роман, рассказ и т.п.), времени и места описываемых событий. Нехудожественные тексты делятся на восемь групп по сфере применения, или функциональному стилю: публицистика (новости и публицистические статьи)[1], учебно–научная (научные и научно–популярные статьи и книги, учебники, лекции и т.д.), официально–деловая (законы, указы, заявления и т.д.), церковно–богословская, рекламная, бытовая (письма, записки и т.д.) и производственно–техническая литература (инструкции, технические паспорта и т.п.). В дополнение к этому используется открытый список типов текстов, например, интервью, инструкция, закон, личное письмо (в настоящее время список содержит более 100 типов). Тематика текстов кодируется списком из 54 категорий, имеющих разную степень дробности: от «экономика» или «политика и общественная жизнь» до «путешествия» или «вооруженные конфликты». В отдельный корпус выделены устные тексты. Они делятся на публичную речь (теле– и радиоинтервью, лекции), непубличную речь (разговоры дома, в магазине, по телефону и т.д.) и речь кино.

Метатекстовая разметка дает возможность поддерживать в корпусе выверенный баланс текстов разных типов. На основе метатекстовой информации можно строить частотные списки на отдельных выборках корпуса и сравнивать их между собой. В частотном словаре тексты были разделены на четыре функциональных стиля: художественная литература, публицистика,

[1] В корпусе проводится различие между функциональным стилем и источником публикации текстов. Например, в газетах представлены как публицистические статьи, так и официально-деловые документы. В журналах также встречаются художественные произведения. Публицистика как функциональный стиль включает в себя новости и публицистические статьи из газет и журналов, информационные тексты, такие как путеводители и рецепты, а также мемуарно-биографическую литературу.

другая нехудожественная литература и устная речь (в объеме подкорпуса непубличной речи). Баланс текстов, представленных в частотном словаре, показан в табл. 1.

Табл. 1. Соотношение текстов разных функциональных стилей в частотном словаре

Функциональный стиль	Доля	Размер подкорпуса, токенов	Размер подкорпуса, орф. слов	Кол–во текстов
Художественная литература	39.04%	45 150 317	35 150 521	2 418
Публицистика	42.21%	48 818 173	39 739 644	27 390
Другая нехудожественная литература,	16.96%	19 618 518	15 478 151	7 495
в т. ч. учебно–научная	11.30%	13 067 152		3 994
официально–деловая	1.62%	1 872 482		1 075
электронная коммуникация	1.49%	1 727 363		133
церковно–богословская	1.44%	1 664 804		488
реклама	0.57%	659 707		1 232
бытовая	0.48%	556 291		439
производственно–техническая	0.26%	295 206		134
Устная непубличная речь	0.88%	1 017 568	758 407	1 005
Другое (в т.ч. смешанный стиль)	0.90%	1 037 468	827 580	61
Итого	100%	115 642 044	91 954 303	38 369

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 частоте 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; примерно в 5% текстов лексико–грамматическая омонимия была снята вручную.

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 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[2].

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

Лемма	Леннгрен	Засорина	Штейнфельд	НКРЯ	Интернет
власть	202	364	138	436	428
думать	609	1094	1058	756	818
загрязнение	69	1	–	15	11
задача	499	421	250	282	292
изучение	193	110	–	75	78
любить	415	632	595	503	650
милый	58	242	135	91	110

Как видим, теоретические рекомендации относительно достаточного размера корпуса в данном случае оказываются не слишком достоверными. Причина этого кроется в том, что исходно допускается 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.[3] В результате частотный список, построенный на основе корпуса, отражает специфику тех текстов, которые попали в него при его составлении.

[2] В точке наибольшего расхождения – ср. частоты глагола *любить* – проявляются различия в методике лемматизации отдельных словоформ, в частности, формы *любимый*.

[3] Кеннет Черч называл эту ситуацию проблемой Норьеги (Church 2000), Адам Килгаррифф - whelk problem (Kilgarriff 1997). Мануэль Норьега – панамский диктатор (1983–1989 гг.). В корпусе, который рассматривал Черч, фамилия диктатора с большой частотой упоминалась в ряде новостных статей 1989 года, посвященных американской военной кампании в Панаме, за пределами этих текстов слово *Норьега* практически не встречалось. *Whelk* – сравнительно редкое английское слово, обозначающее вид моллюска.

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 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.

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, знаки пунктуации и другие символы. 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 *двадцатъ 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, жить–то*), апострофом (*о'кей*), косой чертой *(и/или, км/ч),* как выделять и учитывать обороты (*в течение, невзирая на)* и т. п. В соответствии с использованной моделью токенизации и лемматизации (см. раздел 5 Введения), данный частотный словарь основан на корпусе из 91 982 416 словоупотреблений, включая слова, записанные кириллицей и латинскими буквами, а также числа, записанные римскими и арабскими цифрами. С учетом знаков препинания и другой графики, объем корпуса составляет 115 642 044 токенов (комбинация знаков препинания типа [», --] считается как один токен).

С точки зрения словарного запаса в корпусе современного русского языка содержится 686 566 уникальных типов токенов (лемм, чисел и пунктуации), 1 729 928 отдельных орфографических словоформ, 564 555 кириллических лемм и 70 931 лемм, записанных латиницей. Из кириллических лемм 270 498 встречаются в корпусе более одного раза, 203 185 0 более двух раз, 106 874 – десять раз и более. Десять самых частотных лемм покрывают 16.5% текста, 100 лемм – 37%, 1 000 лемм – 60%, 2 000 лемм – 69%, 10 000 – 85% всех текстов (см. табл. 6.7).

4 Статистические показатели, используемые в словаре

4. 1. Общая частота и ранг

Общая частота характеризует количество 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 39 653 раз в корпусе современного русского языка НКРЯ, то его частота в ipm

составит 137.5, 364.0 и 435.6, соответственно. За единицу вычисления ipm в основной части частотного словаря принято число 92 (сумма орфографических слов корпуса составляет 92 миллиона употреблений). Чтобы примерно оценить абсолютную частоту употреблений некоторого слова в корпусе, надо его частоту F(ipm) умножить на коэффициент 92, например, абсолютная частота существительного *вопрос* составляет 805.8 ipm x 92 = 74 134 употреблений.

На основе общего списка лемм, упорядоченного по частоте, леммам присваивается ранг. Самое употребительное слово – *и* – имеет ранг 1, следующее – *в* – ранг 2 и т.д., редкие слова имеют ранг 10 000 и больше. В отличие от словаря Засориной, где слова с одинаковой частотой имели одинаковый ранг, в настоящем словаре у каждого последующего слова ранг увеличивается на единицу, то есть ранг определяется с помощью простой нумерации общего частотного списка (см. раздел 2). Информация о соотношении рангов лемм и их частоты в ipm указана в приложении к разделу 1, например, леммы, получившие ранг порядка 1 000 (*быстрый, пользоваться, функция*), имеют частоту около 120 ipm. Сведения о покрытии корпуса приводятся в таблице 6.7. раздела 6: так, если для ранга 1 000 указан коэффициент покрытия 0.6094, это означает, что множество лемм с рангом от 1 до 1 000 (первая тысяча лемм) покрывает 60.94% всех словоупотреблений корпуса; 50 000 самых частотных лемм покрывают 93% словоупотреблений корпуса.

Свою систему рангов имеют также словоформы (см. приложение к разделу 4), числа, буквы, знаки препинания (см. раздел 6).

4. 2. Показатель R (range) и коэффициент Жуйана (D)

Задачей частотного словаря является не просто ранжировать слова по их частоте в отдельном корпусе, но и дать материал для определения лексического ядра языка. Необходимо отличать слова, часто встречающиеся во многих текстах, от тех, которые сконцентрированы всего в одном или нескольких текстах корпуса, но употребляются там с большой частотой (ср. примеры со словами *хоббит* или *Норьега* выше). Очевидно, что если бы подбор текстов был другим, то частота слов последнего рода могла бы существенно уменьшиться.

В Чешском национальном корпусе используется понятие средней уменьшенной частоты (ARF, 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, Lyne 1985) используется коэффициент *D*, введенный А. Жуйаном (Juilland's D, см. Juilland et al. 1970; подробный разбор подхода Жуйана и других мер дисперсии см. в Gries 2008).

Коэффициент D отражает равномерность распределения частот в разных сегментах корпуса и вычисляется по следующей формуле:

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

где n – количество сегментов, на которые разбит корпус, μ – средняя частота слова по всему корпусу (т. е. сумма частот в каждом сегменте, поделенная на n), σ – среднее квадратичное отклонение частоты μ на отдельных сегментах.

Для подсчета коэффициента Жуйана корпус разбивается на n равных сегментов (в нашем случае, на 100 частей, размером приблизительно в 90 тыс слов каждый). Тексты в корпусе специально упорядочиваются по функциональным стилям, поэтому тексты одного жанра (например, научные статьи) аккумулируются в пределах небольшого числа сегментов.

Показатель R (range) отражает количество сегментов корпуса, в которых встретилось слово. Коэффициент вариации σ/μ может принимать значения от 0 (в каждом сегменте частоты одинаковы) до 1 (все словоупотребления встречаются только в одном сегменте). Следовательно, значение *D* у слов, встречающихся в большинстве документов, близко к 100, а у слов, часто встречающихся лишь в небольшом числе документов, близко к 0.[4]

Например, союз *но* встретился во всех сегментах корпуса (R=100), его средняя частота – 5381.4 ipm, небольшие колебания частоты в сегментах дают коэффициент Жуйана, равный 97. Существительное *статья* также встречается в 100 сегментах корпуса, средняя частота – 395.0 ipm, но поскольку это слово чрезвычайно часто употребляется в законах и гораздо реже – в художественной литературе, то коэффициент Жуйана равен 76. Слово *конунг* встречается только в девяти сегментах корпуса, средняя частота 10.2 ipm, но при этом оно 916 раз (абсолютная частота) встречается в художественной литературе и всего 3 раза в публицистике и 9 раз в другой нехудожественной литературе, отсюда низкий коэффициент Жуйана – 9.

В основной части словаря (раздел 1) приводится общая частота леммы в ipm, показатели R и D, а также количество документов (текстов), в которых встретилось слово. Каждая из этих мер частоты полезна для определенных целей, но в то же время к ним необходимо подходить с долей осторожности. Так, общая частота может вырастать, если слово активно используется в небольшом количестве текстов. Частота по документам способна это отследить, но она не учитывает того, что некоторые слова встречаются в большом количестве коротких документов (например, новостных сообщений), и это дает им неоправданно высокое значение частоты по документам. Наконец, подсчет сегментов оперирует с объектами одного размера, но не учитывает распределения частоты внутри таких объектов (почти миллион слов в данном словаре). Достаточно, чтобы в сегменте слово встретилось один раз, чтобы его R увеличился на единицу. Например, глагол *пританцовывать* встречается в большинстве сегментов художественной литературы, а также во многих сегментах публицистики (прежде всего, в мемуарах – 47 раз). С учетом того, что глагол также употребляется несколько раз в научных текстах, форумах и устном подкорпусе, он получает высокий R=71. Напротив, для достаточно редких слов количество документов часто близко количеству сегментов.

Мы используем показатель *D* в нашем словаре, так как он является лучшим из известных в настоящее время способов измерить, насколько общеупотребительным является слово, или, напротив, насколько оно специфично для отдельных предметных областей (Lyne 1986). Например, прилагательные *преподобный, геологический* и *внимательный* имеют в НКРЯ примерно равную частоту (около 25 ipm), но при этом коэффициент *D* у *преподобный* – 46, *геологический* – 78, а у *внимательный* – 97, что означает, что последнее слово значимо для большего числа предметных областей и (при прочих равных условиях) имеет большие шансы на место в неспециализированном словаре.

Низкий D в сочетании с высоким R (range) служит также своеобразным предупреждением о том, что частота слова в словаре завышена: в одном или нескольких текстах корпуса это слово является темой. Например, такова ситуация со словом *якорь* в нашем словаре: это имя достаточно равномерно распределено по всему корпусу (R=91) и в целом не слишком частотно, но в одной только «Книге о якорях» оно встречается более 400 раз (D=28).

Тем не менее, ранжирование словника с использованием коэффициента D представляет проблему, и это связано с тем, что соотношение значений *D* у слов с разной абсолютной частотой неочевидно. Определенный компромисс предлагался в словаре Л. Леннгрена: в нем

[4] Здесь мы следуем методу презентации в Leech et al. 2001. В оригинале коэффициент считается без умножения на 100. В настоящем словаре значение коэффициента округляется до целого.

частотный список был отсортирован по значению произведения коэффициента *D* на среднюю частоту слова (т. н. модифицированная частота). Однако, в связи с тем, что теоретический статус этого произведения неясен, мы не считали целесообразным сортировать наш словарь по нему.

4. 3. Коэффициент логарифмического правдоподобия LL–score (значимая лексика)

Наиболее частотные служебные слова приблизительно равномерно употребляются в текстах разных стилей и жанров. В то же время, частота слов *процесс* и *теория* в научных публикациях значительно превышает их частоту во всех остальных текстах корпуса. Аналогичным образом, слова *ну, да, вот, пожалуйста* употребляются чаще в устной речи, а слова *сказать, спросить, локоть, снег* – в художественной литературе. Сравнивая частоты слов в разных подкорпусах, можно получить списки значимой лексики[5] для того или иного функционального стиля.

В качестве метрики сравнения используется 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 G^2 (LL–score) можно вычислить как:

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

Здесь *a*, *b*, *c*, *d* – наблюдаемые величины, а *E1* и *E2* – ожидаемый показатель в сравниваемых подкорпусах (см. Rayson & Garside 2000).

Этот показатель учитывает как относительную частоту (во сколько раз чаще слово встречается в подкорпусе по сравнению с остальными текстами), так и абсолютную частоту в корпусе и подкорпусе. Последнее обстоятельство важно, поскольку значимость того, что слово встретилось в подкорпусе в 10 раз чаще чем в основном корпусе, зависит от того, имеем ли 5 или 500 вхождений этого слова в подкорпус. В первом случае она может быть связана со случайными флуктуациями, во втором эти данные статистически значимы.

Достоинством критерия правдоподобия является и то, что возможна статистическая оценка значимости различия частот в подкорпусе и остальном корпусе. Если этот показатель превышает 15.31, с вероятностью более 99% можно отвергнуть гипотезу, что разница в частоте случайна и она не обусловлена существенными различиями в составе корпуса (Rayson & Garside, 2000).

Матрица, приведенная ниже, иллюстрирует примеры подсчета по этой формуле. Во всех четырех случаях отношение нормализованных частот (в ipm) в подкорпусе и корпусе одинаково (полтора к одному). В первом и третьем примерах одинаковая нормализованная частота (15 ipm в подкорпусе и 10ipm в корпусе), но размер подкорпуса (a) и корпуса (a+b) в третьем примере в десять раз меньше, чем в первом. В первом и втором примерах подкорпус и корпус одного размера, но количество употреблений во втором примере в десять раз меньше.

[5] В Шайкевич 2003: 17-19 лексика подобного рода называется «лексическими маркерами».

Наконец, в первом и четвертом примерах при одинаковых значениях ipm и общем размере корпуса отличаются размеры подкорпусов (они соотносятся как 10:1).

Табл. 3. Зависимость LL–score от частоты леммы и размеров корпусов

	подкорпус	корпус	подкорпус	корпус	подкорп.	корпус	подкорп.	корпус
Частота абс.	300	1000	30	100	30	100	30	1000
Размер	20000000	100000000	20000000	100000000	2000000	10000000	2000000	100000000
Частота ipm	15	10	1.5	1	15	10	15	10
E1	200		20		20		20	
E2	800		80		80		980	
LL	56.34		5.63		5.63		4.43	

В соответствии с критерием правдоподобия первый пример более значим (300 фактов употребления дают большую статистическую значимость, невзирая на относительные частоты) и только первый пример является статистически значимым (значение коэффициента превышает 15.31).

Все же (как и в случае с абсолютной частотой или коэффициентом D) не стоит абсолютизировать важность конкретного значения этого критерия. Корпуса далеки от совершенства, показатели могут отражать случайные параметры их создания, а не устойчивые параметры языка (Kilgarriff, 2005). Например, язык 1950–60–х годов в корпусе намного лучше отражен жанром художественной литературы (в настоящий момент доступно относительно небольшое количество нехудожественных текстов этого периода), поэтому «статистически важными» лексическими маркерами нехудожественных текстов этого периода в НКРЯ являются *куст* и *землянин* (наряду со словами *советский, коммунистический, товарищ*, которые действительно отражают специфику этого периода).

5 Принципы создания словника

5.1 Размер словаря

Хорошо известно, что распределение частот лексических единиц крайне неравномерно: очень небольшое количество слов встречаются 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.

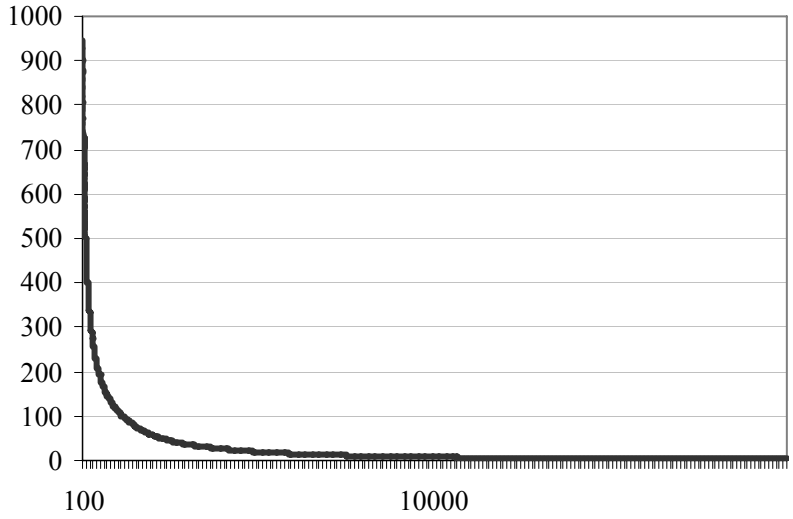

Рис. 1: Зависимость частоты от ранга (закон Ципфа).

Чем дальше от начала списка, тем менее предсказуемой становится частота конкретного слова и тем больше она зависит от текстового наполнения корпуса. Так, по данным НКРЯ слова *неумолимо* и *подвох* входят в число 20 000 самых частотных слов, а слова *изворотливый* и *раскуривать* находятся за пределами списка первых 30 000, что не вполне соответствует интуиции.

В частотных словарях принято вводить порог, ограничивающий список слов по частоте употребления. Величина порога зависит от полиграфических возможностей издания, а также от размера корпуса, на котором строится словник. Практика создания частотных словарей на материале 100–миллионных корпусов обычно ограничивает частотные списки словами с частотой около 5 употреблений на миллион слов (ipm), что для корпуса современного русского языка НКРЯ дает словник около 13 000 слов (самые редкие слова встречаются в корпусе около 460 раз). Такой объем, безусловно, обеспечивает представление о ядерной лексике и вполне достаточен, например, для изучения русского языка как иностранного. Тем не менее, не следует забывать, что частотные словари должны отвечать и на другой вопрос – какие именно единицы входят в словарный фонд языка, и в частности, что находится за пределами лексического ядра? Очевидно, что указанный объем словника явно мал для задач лексикографии и компьютерной лингвистики и не дает материала для сопоставления настоящего словаря с другими частотными словарями русского языка, в частности, с 40–тысячным словарем Засориной. В связи с этим частотный порог в нашем словаре был снижен до 2.6 ipm для рангового списка (раздел 2, 20 000 лемм) и до 0.4 ipm для алфавитного списка (раздел 1, около 50 000 лемм, самые редкие слова встречаются в корпусе 33 раза).

5.2 Лемматизация и частеречная аннотация

При подготовке словаря каждой словоформе корпуса был приписана лемма (исходная форма) и часть речи. Разметка была проведена по пословному принципу: устойчивые обороты, составные предлоги и другие неоднословные лексические единицы (ср. *Новый год, в течение, тем не менее, друг друга*) словаре отсутствуют; составляющие их орфографические слова учтены по отдельности (при леммах *новый, год, в, течение*, *то, не*, *менее, друг*).

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 анализ и сбор данных происходит в автоматическом режиме.

Каждой лемме приписывается информация о части речи. Выделяются следующие классы:

s — существительные (*яблоня, лошадь, корпус, вечность*),
a — прилагательные (*коричневый, таинственный, морской*),
num — числительные (*четыре, десять, много*),
anum — числительное–прилагательные (*один, седьмой, восьмидесятый*)
v — глаголы (*пользоваться, обрабатывать*),
adv — наречие (*сгоряча, очень*); в этот класс включены также предикативы (*жаль, хорошо, пора*) и вводные слова (*кстати, по–моему*),
spro — местоимения–существительные (*она, что*),
apro — местоимения–прилагательные (*который, твой*),
advpro — местоименные наречия (*где, потому*); в этот класс включены также местоимения–предикативы (*некого, нечего*),
pr — предлоги (*под, напротив*),
conj — союзы (*и, чтобы*),
part — частицы (*бы, же, пусть*),
intj — междометия (*увы, батюшки*),
init — инициалы (*А., Вл.*)[6].

При разночтениях в словарных источниках, связанных с классификацией слов по частям речи, морфологический стандарт НКРЯ придерживается в основном трактовки Грамматического словаря А.А. Зализняка. Два класса, отсутствующие в (Зализняк 1977) – местоименные наречия (ср. *где, здесь, так, как, куда–нибудь, нисколько, везде*; в грамматическом словаре считаются наречиями) и инициалы (включая однобуквенные типа *А.*и неоднобуквенные типа *Вл., Вяч.*). В соответствии с трактовкой Грамматического словаря, слова *остальной* и *другой* относятся к местоимениям–прилагательным, все употребления слова *вот* – к частицам. Формы причастий входят в парадигму глагола. Возвратные и невозвратные глаголы, глаголы совершенного и несовершенного вида считаются отдельными единицами словаря.

Слова, записанные через дефис, лемматизируются или как одна единица (ср. *кое–как, гран–при, жилищно–строительный, жар–птица*), или каждой части приписывается собственная лемма (ср. *город–спутник, член–корреспондент*). Программа автоматического разбора распознает как одну лемму только те слова с дефисами, которые были включены в ее словарь (список этих слов во многом совпадает со списком Грамматического словаря А.А. Зализняка). Написания через косую черту (ср. *км/ч*) аннотируются как две леммы и учитываются по отдельности (ср. *км, ч*). Из написаний с апострофом лемматизированы как одна единица и включены в словарь только слово *о'кей* и его орфографический вариант *о'кэй*.

Заимствованные слова, записанные кириллицей, учитываются в том случае, если в корпусе имеется достаточно данных, что они подчинились русской системе словоизменения (так, включено слово *шоп*, имеющее употребления *в шопе*, *в шопах* и т.д., но не включено слово

[6] Программы Mystem и Dialing адаптированы к морфологическому стандарту НКРЯ, однако допускают расхождения в частеречной аннотации отдельных слов. Для достижения лучших результатов разрешения омонимии разборы были стандартизованы, а некоторые частеречные классы НКРЯ укрупнены, см. классы наречий и местоименных наречий.

лимитед, встечающееся лишь в названиях предприятий, построенных по западному образцу (ср. *«Омега Лимитед»*).

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 (*ходют*). 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 и др. 2007), а затем выверены вручную.

5.3 Разрешение лексико–грамматической неоднозначности

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 (Čermak et al. 2004).

Первоначально корпус современного русского языка был размечен двумя программами морфологического анализа, которые каждой словоформе приписывали все потенциальное множество разборов. Особая часть корпуса (объемом 4,8 млн. словоупотреблений, что составляет около 5% всех текстов) была размечена программой Dialing (Сокирко 2004); затем в этом подкорпусе была вручную снята грамматическая омонимия. Основной массив корпуса был размечен морфологическим анализатором Mystem (Сегалович & Маслов 1998). Здесь неоднозначность в лексико–грамматической разметке была разрешена с помощью компьютерной программы, использующей модель триграмм и обучаемой на вышеупомянутом тренировочном подкорпусе со снятой вручную омонимией (разработчики программы дизамбигуации А.В. Сокирко, А.И. Зобнин и др., ООО «Яндекс», см. Сокирко & Толдова 2005). Точность определения леммы и части речи автоматическим способом составила 93.81%.[7] Поскольку автоматическое разрешение омонимии допускает определенную погрешность, омонимы, входящие в первые 20 тысяч частотных слов, подверглись выборочной ручной проверке.

Как правило, сочетание словоклассифицирующих грамматических тегов «часть речи + род» позволяет однозначно предсказать лемму у неоднозначной словоформы. Тем не менее, остается небольшое число вариантов, для которых алгоритм программы не смог обеспечить выбор леммы, а именно:

[7] Учитывались только неоднозначные словоформы; программа обучалась на комбинации грамматических тегов «часть речи + род + число + падеж». Для словоформ за пределами списка 3 000 наиболее частотных этот показатель составил 93.07%. Доля правильно определенных словоформ по отношению ко всем словоформам корпуса составила 97-98%.

1) леммы с двумя вариантами исходной формы, у которых словари традиционно не признают разницы в значении, ср. *достичь–достигнуть, постичь–постигнуть, стыть–стынуть, остыть–остынуть, застыть–застынуть*. В словаре указан лишь один вариант леммы (ср. *достигнуть, остынуть*).
2) слова, имеющие варианты исходной формы, с небольшой разницей в значении, ср. *гастроли–гастроль, доспехи–доспех, расценка–расценок, малолеток–малолетка, шпрот–шпрота, овсяный–овсяной, договорный–договорной, святой–святый, валовый–валовой, занятый–занятой, запасный–запасной*. Здесь был применен принцип «основной леммы»: один из вариантов был признан основным, под ним учтена частотность всех словоформ, входящих в эту лемму. Формы, не входящие в парадигму этой леммы, были учтены при подсчетах частоты другой леммы. Например, у вариантов *доспехи–доспех* основной была признана лемма множественного числа *доспехи*, формы единственного числа были учтены под леммой *доспех*; у вариантов *святой–святый* все словоформы, кроме *святый*, были учтены под леммой *святой.*
3) имена pluralia tantum, разошедшиеся в значении с леммой, имеющей исходную форму в единственном числе, ср. *плавки–плавка, духи–дух, часы–час,* а также формы сравнительной степени типа *раньше, выше, ниже, дальше, далее*, разошедшиеся в значении с соответствующими наречиями *рано, высоко, низко, далеко.* В данном случае подсчет частот для обоих вариантов был произведен приблизительно на основе распределения данных в выборках корпуса со снятой вручную омонимией.

5.4 Дополнительные соглашения

Лексические омонимы типа *лук*[1] *– лук*[2], *повезти*[1] *– повезти*[2], *вера – Вера*, т. е. слова одной части речи, с одинаковой исходной формой (именем леммы), 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*, *падеж – падёж*).

Омонимичные леммы, принадлежащие к разным частям речи, приводятся отдельно:

Табл. 4. Частотные данные для имени существительного печь *и глагола* печь

Lemma	PoS	F(ipm)	R	D	Doc
печь	s	32.6	100	93	952
печь	v	8.7	95	93	511

Орфографическая норма. Написание слов дается в том виде, как они встретились в корпусе. Случайные опечатки, как правило, имеют низкую частотность, и следовательно, отсутствуют в списках частотного словаря. В то же время, в словаре встречаются слова типа *дагерротип, заграницей, брэнд*, в отношении которых орфографическая норма менялась с течением времени, или клишированные искажения типа *седни* (вместо сегодня), *сичас* (вместо сейчас), *вобщем* (вместо нормативного *в общем*), написание которых отражает тенденции современной неформальной речи в сети Интернет.

Сокращения, которые по правилам русского языка записываются со строчной буквы и с точкой на конце, расшифровываются: например, леммами слов *рис.* и *тел.* считаются, соответственно, *рисунок* и *телефон.* Сокращения, допускающие несколько разборов (например, *стр. – страница*, *строение* и др.), не учитываются в словаре в связи с ограничениями, налагаемыми технологией его подготовки.

В корпусе присутствует некоторое количество написаний типа *преж–ние, очевидно–стью, за'мок*, в которых отражаются следы переноса части слова на новую строку или места ударения в оригинальных бумажных или электронных версиях текстов. Другой источник фрагментированных слов – сокращения вида *отд–ние* (ср. *отделение*). Данные написания не были учтены при составлении словаря.

По техническим причинам, связанным с автоматической обработкой словника корпуса, не учитываются особенности написания слов с прописной VS строчной буквы. Все слова в разделах словаря, посвященных нарицательной лексике, записаны со строчной буквы (в том числе притяжательные прилагательные типа *люсин/Люсин, митин/Митин*), слова в разделе 8 «Имена собственные» – с прописной буквы, за исключением ряда стандартных сокращенных написаний типа *км, кВч*. Варианты типа *км* и *КМ* приведены под общей леммой.

6 Структура словаря

Словарь состоит из следующих разделов:

1. Алфавитный список лемм (общая лексика)
2. Ранговый список лемм (общая лексика)
3. Жанровые особенности (общая лексика)
3.1.а. Частотный словарь художественной литературы
3.1.б. Значимая лексика художественной литературы
3.2.а. Частотный словарь публицистики
3.2.б. Значимая лексика публицистики
3.3.а. Частотный словарь другой нехудожественной литературы
3.3.б. Значимая лексика другой нехудожественной литературы
3.4.а. Частотный словарь устной речи
3.4.б. Значимая лексика устной речи
4. Алфавитный список словоформ
5. Ранговые списки частей речи
5.1. Имена существительные
5.2. Глаголы
5.3. Имена прилагательные
5.4. Наречия и предикативы
5.5. Местоимения (местоимения–существительные, прилагательные, наречия, предикативы)
5.6. Имена числительные
5.7. Служебные части речи
6. Вспомогательные таблицы
– данные о частоте частеречных классов и другая статистическая информация
7. Алфавитный список имен собственных и аббревиатур

В **алфавитном списке лемм** (раздел 1) приводится имя леммы, часть речи PoS, общая частота леммы F(ipm), показатель R (range), коэффициент вариации D и количество документов (текстов) Doc, в которых она встретилась. Алфавитный список включает около 50 000 наиболее частотных лемм общей (нарицательной) лексики. Если нарицательное имя употребляется в корпусе также как имя собственное (см. *вера – Вера*, *кулик – Кулик*), оно снабжается пометой (*). Завершает раздел 1 **таблица рангов лемм**, в которой указано, какую частоту в ipm имеют

словоформы, занимающие 100–е, 1000–е, 100000–е и т.п. место в списке, упорядоченном по частоте употребления.

Табл. 5. Фрагмент раздела 1 (алфавитный список лемм)

Lemma	PoS	F(ipm)	R	D	Doc
абстрагирование	s	0.5	15	63	22
абстрагировать	v	0.4	18	72	25
абстрагироваться	v	1.0	51	85	76
абстрактно	adv	0.7	41	84	54

В **ранговом списке лемм**, упорядоченном по частоте (раздел 2), указываются частотный ранг Rank, имя леммы, часть речи PoS, общая частота F(ipm) и распределение частоты по временным интервалам (1950–1969 годы, 1970–1989 годы, 1990–2007 годы) в подкорпусах художественной литературы и публицистики[8]. Ранговый список включает 20 000 самых употребительных лемм общей лексики.

Табл. 6. Фрагмент раздела 2 (ранговый список лемм)

			худ. литература			публицистика		
Lemma	PoS	F(ipm)	1950-1960	1970-1980	1990-2000	1950-1960	1970-1980	1990-2000
отставка	s	32.5	6.4	11.0	15.7	22.2	23.9	64.5
перестройка	s	32.5	2.1	4.0	16.0	10.4	74.5	52.1

Табл. 7. Сравнительные размеры подкорпусов художественной литературы и публицистики, разбитых по двадцатилетиям

Функциональный стиль		1950–1969	1970–1989	1990–2007
Художественная литература:	орф.слов	5 642 070	7 818 865	21 756 323
	текстов	309	585	1 524
Публицистика:	орф.слов	674 566	2 725 968	34 950 394
	текстов	509	623	26 264

Несмотря на то, что большая часть корпуса состоит из текстов 1990–2000–х годов (см. табл. 7), частотное распределение по годам дает возможность приблизительно оценить микродиахронию – как менялась частота слов в отдельных жанрах за последние 60 лет. Пользователи словаря, однако, должны помнить, что подкорпусы несопоставимы по размеру и отличаются по составу жанров публицистики и тематике художественной литературы, что может отражаться в частотных данных.

[8] Категоризация текстов с большим интервалом времени создания проводилась по поздней дате, например, текст, написанный в 1975-2003 годах, был включен в подкорпус 1900-2000-х годов.

Чтобы найти место некоторого слова в частотном словнике, нужно обратиться к алфавитному списку (раздел 1). Если частота искомого слова больше или равна 2.6 ipm, оно присутствует в ранговом списке лемм.

Частотные словари функциональных стилей (раздел 3) составлены на основе подкорпусов художественной литературы, публицистики, другой нехудожественной литературы и устной непубличной речи. В список включены 5 000 самых частотных лемм этих подкорпусов. При каждой лемме указывается часть речи и частота F(ipm). В приложении даны списки наиболее частотных лемм соответствующего подкорпуса, упорядоченные по рангу.

Табл. 8. Фрагмент раздела 3.4 (жанровые особенности: частотный словарь художественной литературы)

Lemma	PoS	F(sp)
американец	s	22.7
американский	a	26.5

Отдельно приводятся **словари значимой лексики** – список наиболее типичных слов для каждого типа текстов был выделен на основе сравнения частоты лемм в таких текстах и в остальном корпусе (см. п. 4). В частности, значимыми для устного подкорпуса[9] оказываются слова *ну, да, вот, там, угу* и др. – они появляются в разговоре в десятки раз чаще, чем в подготовленной письменной речи. В словарях значимой лексики при каждой лемме приводятся часть речи, общая частота во всем корпусе F(all) в ipm, частота в подкорпусе данного функционального стиля, также в ipm, и коэффициент правдоподобия LL–score.

Табл. 9. Фрагмент раздела 3.4 (значимая лексика устной речи)

Lemma	PoS	F(all)	F(sp)	LL
ну	part	1114.6	17208.0	50672
да	part	787.5	11847.0	34394
вот	part	1785.1	15698.6	32662

Алфавитный список словоформ (раздел 4) включает все словоформы корпуса с частотой выше 5 ipm, представляющих как общую лексику, так и имена собственные (объем списка около 20 тысяч). Для каждой словоформы указана ее частота во всем корпусе. Данные для словоформ, записанных с помощью прописных и строчных букв, приводятся отдельно, различается также написание через *е* и *ё.* Таким образом, в таблице представлено пять единиц:

Табл. 10. Фрагмент раздела 4 (алфавитный список словоформ)

Word	F(ipm)
все	3504.1
Все	631.5
ВСЕ	5.5
всё	276.9
Всё	45.7

[9] Как оказалось, значимая лексика устной публичной речи, особенно интервью и лекций, которые обычно полностью или частично продуманы и подготовлены говорящим заранее, практически ничем не отличается от значимой лексики публицистики. В связи с этим частотный словарь устной речи и словарь значимой лексики устной речи был подготовлен на основе подкорпуса устной непубличной речи, куда вошли бытовые разговоры, микродиалоги в магазине, пересказы снов, споры и т. п.

В приложении приводятся **данные о наиболее частотных словоформах** и **таблица рангов словоформ**, в которой указано, какую частоту в ipm имеют словоформы, занимающие 100–е, 1000–е, 100000–е и т.п. место в списке, упорядоченном по частоте употребления.

В разделе 5 (**ранговые списки частей речи**) частотный список лемм разбит на восемь подсписков: имена существительные, глаголы, имена прилагательные, наречия (в т. ч. предикативы и вводные слова), местоимения, числительные, служебные части речи (предлоги, союзы, частицы, междометия). Для каждой леммы указана ее общая частота F(ipm) и ранг (порядковый номер) в общем списке Rank. Каждый список содержит 1 тысячу наиболее частотных лемм. В приложении приводится частотный список чисел, записанных цифрами.

Табл. 11. Фрагмент раздела 5.7 (ранговые списки частей речи: служебные части речи)

Lemma	PoS	F(sp)
прежде	pr	147.3
зато	conj	134.9

Вспомогательные таблицы (раздел 6) включают в себя данные о частоте частеречных классов, букв русского алфавита и их сочетаний, знаков препинания, а также информацию о покрытии текста лексемами и о длине текстов и словоформ. В разделе 6.1 приводится абсолютная частота F(abs) и доля употреблений (%) частеречных классов по данным подкорпуса с ручным разрешением омонимии и подкорпуса с автоматическим разрешением омонимии. В разделах 6.2–6.5 сообщается о частоте букв русского алфавита, а также двух–, трех– и четырехбуквенных сочетаний. Данные упорядочены по алфавиту; приводится абсолютная частота F(abs) и ранг элемента Rank. Раздел 6.6 посвящен частоте знаков препинания (упорядочено по рангу, также приводится абсолютная частота и ранг элемента).

Раздел 6.7 содержит данные о покрытии корпуса: для каждого ранга леммы (Rank) приводится накопленная частота (Coverage). Например, из таблицы следует, что наиболее употребительная лемма (с рангом 1) покрывает 3.6% текстов, т. е. 3.6% словоупотреблений в корпусе приходится на союз *и*, леммы с рангом 1–2 вместе покрывают 6.7% текста, леммы с рангом 1–10 вместе покрывают 16.6% текстов, а 93% словоупотреблений корпуса приходится на леммы с рангом 1–50000.

В разделе 6.8 приводятся данные о длине текстов. Тексты разделены на категории 1–100 слов, 101–200 слов и т. д., и для каждой категории указано количество текстов в подкорпусе художественной литературы NT(im), публицистики NT(n) и другой нехудожественной литературы NT(nf). Данные для подкорпуса устной речи не приводятся, так как в нем деление на тексты имеет условный характер. В разделе 6.9 указаны данные о длине словоформ: длина (L), пример (Example), количество разных словоформ с данной длиной (N) и их совокупная частота в ipm (F) для всего корпуса (all) и для подкорпусов художественной литературы (im), публицистики (n), другой нехудожественной литературы (nf) и устной речи (sp). График иллюстрирует сравнительную частоту словоформ разной длины в подкорпусах.

Завершает словарь **алфавитный список имен собственных и аббревиатур** (раздел 7). Имена собственные отделены от основной части словника, так как образуют значительно менее стабильную в статистическом отношении группу, а их частота 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. Имена героев художественных произведений могут

повторяться достаточно часто в пределах одного текста, поэтому высокий показатель частоты ipm может быть обманчив. Критерии равномерности распределения имен на массиве корпуса (количество текстов Doc, коэффициенты R и D) для имен собственных приобретают большую ценность. Частотный список имен собственных также сильно зависит от выбора текстов для корпуса. Чтобы хотя бы в некоторой степени ослабить влияние этого фактора, в данный список мы включили имена, которые встретились в корпусе не менее 150 раз (1.6 ipm) и не менее чем в 50 текстах корпуса.

Для получения списка имен собственных и аббревиатур из конкорданса корпуса были выделены имена существительные, написание которых в текстах с большой буквы превышало 90–процентный порог, ср. *Россия, Смирнов, ГРЭС, МИД, КЗоТ*. В этот же список вошли употребительные единицы измерения, такие как *мл, МПа, кВт* и т.п. Вместе с тем, имена *Бог, Аллах, Будда*, названия священных книг и религиозных праздников, названия средств массовой информации, транспорта и т.п., восходящие к нарицательным именам (*«Известия», «Автопилот», «Варяг»*), приводятся в разделах общей лексики (1–5). Прилагательные с большой долей написаний с заглавной буквы, например, *Христов, Петин, Костромской*, также отнесены к общей лексике.

В раздел 7 включена ядерная часть списка, насчитывающая 2 500 наиболее употребительных единиц. Как и в разделе 1, для каждого существительного приводится общая частота F(ipm), показатели R и D и количество документов Doc.

Табл. 12. Фрагмент раздела 7 (алфавитный список имен собственных и аббревиатур)

Lemma	F(ipm)	R	D	Doc
Алексеев	9.1	72	88	372
Алексеевич	52.4	99	67	522
Алексеевна	12.0	90	87	275
Алексей	115.9	100	91	3387
Алексий	11.3	57	82	305

Если существительное употребляется в корпусе преимущественно как имя собственное, но наряду с этим имеет нарицательное употребление (см. *Василек – василек, Майя – майя*), оно снабжается пометой (*). Написание имен собственных и аббревиатур отражает наиболее часто встречающийся в корпусе вариант (так, из возможных вариантов написания *АВТОВАЗ, АвтоВАЗ, Автоваз* выбран наиболее употребительный *АвтоВАЗ*, для имени/фамилии *Мур* и аббревиатуры *МУР* приводится более частотный вариант *Мур*). Мужские и женские фамилии на *–ов(а), –ев(а), –ин(а), –ын(а), –ский/–ская, –цкий/–цкая* и т.п. приводятся только в мужском варианте, частоты словоформ мужской и женской парадигмы объединены. В приложении приводится алфавитный список **инициалов**, для которых указана частота F(ipm).

***

Авторы пользуются случаем выразить благодарность Институту русского языка им. В.В. Виноградова РАН (Москва), Университету Тромсе (Universitetet i Tromsø, Норвегия), Университету Лидса (University of Leeds, Великобритания), а также Фонду им. Гумбольдта за предоставленную возможность работы над проектом частотного словаря. Наша самая искренняя признательность команде Национального корпуса русского языка, в частности, всем,

кто собирал и обрабатывал материалы основного и устного корпуса, а также создавал подкорпус с ручным разрешением грамматической омонимии – этот словарь появился на свет благодаря их упорному кропотливому труду. Наша особая благодарность Е.В. Рахилиной, человеку, который превратил абстрактную идею Национального корпуса в реальный проект, а также А.Я. Шайкевичу за внимательную оценку первой версии словаря. Мы признательны за поддержку и высказанные ценные замечания Е.А. Гришиной, Ю.Л. Кузнецовой, О.А. Митрофановой, Е.В. Падучевой, В.А. Плунгяну, Д.В. Сичинаве, С.Ю. Толдовой и многими другими коллегами, с которыми мы имели счастливую возможность обсудить концепцию словаря, его наполнение и текст введения. Благодарим также Г.К. и Д.К. Бронниковых, Б.П. Кобрицова, Г.И. Кустову, С.О. Савчук, Д.В. Сичинаву, О.М. Урюпину за техническую помощь и советы при подготовке материалов к словарю, сотрудников компании «Яндекс» А.А. Аброскина, Н.В. Григорьева, С.В. Давыдова, А.И. Зобнина и А.В. Сокирко, которые разрабатывали и совершенствовали программы разметки и дизамбигуации корпуса, а также А.В. Санникова, осуществившего электронную публикацию словаря. На завершающем этапе неоценимую помощь в проверке материала оказала редактор издания И.В. Нечаева.

Словарь создан в рамках Федеральной целевой программы Федерального агентства по образованию РФ «Русский язык» (Госконтракт П66). Настоящая публикация осуществлена при финансовой поддержке Российского гуманитарного научного фонда в рамках проекта «Образ России в современном мире».

Электронная версия словаря опубликована на сайте Института русского языка им. В.В.Виноградова РАН (http://dict.ruslang.ru).

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.

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.* М.: Русский язык.

Лённгрен 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., Сичинава Д.В., Кобрицов Б.П. (2007). Автоматизация построения словаря на материале массива несловарных словоформ // Браславский П.И. (ред.), Интернет–математика 2007. Екатеринбург: Изд–во Урал. ун–та. С. 118–125.

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

НКРЯ 2006–2008: *Национальный корпус русского языка 2006–2008.* СПб.: Нестор–История, 2009.

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 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.В., Маслов М.Ю. (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. (2004). Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2004». М. http://www.dialog-21.ru/Archive/2004/Sokirko.pdf.

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.: Яndex. С. 80–94.

Степанова Е.М. (1970). *Частотный словарь общенаучной лексики*. М.: Изд–во МГУ.

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

Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. (2008). *Статистический словарь языка русской газеты* (*1990–е годы*). 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.

Čermák, František, Michal Křen et al. (2004). Frekvenční slovník češtiny. Praha: NLN.

Church, Kenneth W. (2000). Empirical estimates of adaptation: the chance of two Noriegas is closer to *p/2* than p^2 // Proceedings of the 17th conference on Computational linguistics. Saarbrucken, Germany, 2000. 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. C. 1327–1333.

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

Davies, Mark & Dee Gardner (2010). *A Frequency Dictionary of American English: Word Sketches, Collocates, and Thematic Lists*. London–N.Y.: Routledge. http://www.wordfrequency.info/

Gries, Stefan Th. (2008). Dispersions and adjusted frequencies in corpora // International Journal of Corpus Linguistics 13, 4. P. 403–437.

Josselson Harry H. (1953). *The Russian word count and frequency analysis of grammatical categories of Standard Literary Russian*. Detroit: Wayne University Press.

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.

Kilgarriff, Adam (2005). Language is never ever ever random // Corpus Linguistics and Linguistic Theory 1 (2): 263–276. http://www.kilgarriff.co.uk/Publications/2005-K-lineer.pdf

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

Lyne, Anthony A. (1986). In Praise of Juilland's 'D'; a contribution to the empirical evaluation of various measures of dispersion applied to word frequencies // Ch. Muller (ed.) Methodes quantitatives et informatiques dans l'etude des textes. Geneve–Paris. P. 588–595.

Lyne, Anthony A. (1985). The vocabulary of French business correspondence: word frequencies, collocations and problems of lexicometric method. Genève: Slatkine, Paris: Champion. (Travaux de linguistique quantitative, 23).

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. P. 63–98. http://wackybook.sslmit.unibo.it.

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