ЧАСТОТНЫЙ СЛОВАРЬ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: КОНЦЕПЦИЯ И ТЕХНОЛОГИЯ СОЗДАНИЯ

FREQUENCY DICTIONARY OF THE RUSSIAN NATIONAL CORPUS: PRINCIPLES AND TECHNOLOGY

Ляшевская О.Н. *(olesar@mail.ru), Институт русского языка им. В.В. Виноградова РАН*
Шаров С.А. *(s.sharoff@leeds.ac.uk), Университет Лидса, Великобритания*

Словарь содержит представительный базовый словник современного русского языка (2-я половина XX – начало XXI вв.), снабженный информацией о частотности употребления, статистическом распределении по текстам и жанрам, по времени создания текстов. Словарь основан на текстах Национального корпуса русского языка объемом 100 млн. словоупотреблении.

1. Введение

Для русского языка было разработано несколько частотных словарей. Пионером был словарь Г. Йоссельсона, изданный в 1953 году в Детройте на материале языка по преимуществу дореволюционной России. Словари Э.А. Штейнфельд (1963), Л.Н. Засориной (1977), Л. Леннгрена (1993) и др. 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, а слово *расческа* отсутствует. Существуют также специализированные словари, в частности, словарь Е.М. Степановой (1976), посвященный общенаучной лексике. 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.

Новый частотный словарь – универсальный. Несмотря на то, что последний его прямой предшественник был выпущен 15 лет назад (Леннгрен 1993), очевидно, что за это время изменилось многое – как сам язык, так и технология подготовки частотных словарей. Наш словарь призван представить статистическую картину современного словоупотребления (1950-2005 г.), заполнив, в частности, лакуну последних двух десятилетий, а также показать изменения, произошедшие в языке с 1950 года.

Словарь базируется на 100-миллионном корпусе, в то время как предыдущие словари опирались на материал объемом от 400 тыс. до 1 млн. словоупотреблений. Национальный корпус (www.ruscorpora.ru, НКРЯ 2005) более представителен по охвату материала, так как содержит сбалансированную коллекцию текстов разных типов, жанров и стилей, в том числе и тексты русского зарубежья. Распределение текстов в подкорпусе современного русского языка (с 1950 года) по функциональным стилям показано в таблице 1. Тексты нехудожественной литературы относятся к более чем 50 предметным областям (экономика и финансы, право, путешествия и др.), а их типология варьируется от законов и научных статей до интервью, инструкций и объявлений (всего более 100 типов). Художественные тексты включают романы, повести, рассказы, очерки, пьесы, сказки, эссе, литературные письма и др.

Художественная литература	36%
Публицистика	42%
Прочая нехудожественная литература	17%
Устная литература	5%

Таблица 1. Функциональные стили подкорпуса современного русского языка

Большой размер и стилистическая сбалансированность корпуса являются предпосылкой того, что он будет давать надежные статистические результаты для наиболее частотных слов: так, состав первых 20 000 элементов не будет существенно меняться, если, сохранив пропорцию, заменить данные тексты другими или сравнить несколько подвыборок корпуса. Это показывает опыт составления частотных словарей других 100-миллионных

национальных корпусов, таких как британский, чешский (Leech et al. 2001, Čermák & Křen 2004), а также корпуса испанского языка (Davies 2005). Естественно, что частотный словарь НКРЯ во многом, и в технологических вопросах, и содержательно, ориентируется на эти образцы.

2. Размер корпуса и надежность выборки

Существующие частотные словари для русского языка были построены на сравнительно небольших корпусах: 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.

Лемма	Леннгрен	Засорина	Штейнф.	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	110

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

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 (кор-

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

пус 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 *Норьега* или *хоббит*, и которые случайно оказались в той или иной позиции частотного списка. Так 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) 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:

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

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. 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 употребление на миллион слов), но при этом коэффициент D у *специфический* - 66, *сырье* - 18, а у *жуткий* - 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.

3. Структура словаря

Концепция словаря предполагает издание «бумажной» версии с сопутствующим ей электронным вариантом, представляющим частотный словарь в более полном объеме. Словарная часть содержит следующие разделы:

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 живой устной речи,
    словарь значимой лексики живой устной речи
- алфавитный список словоформ

II. 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 электронного издания включает 60 000 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 по следующей формуле (Rayson & Garside 2000):

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

Словари значимой лексики для разных функциональных стилей включают по 500 лемм.

Алфавитный список словоформ включает все словоформы корпуса с частотой выше 0.1 ipm (всего около 15 тыс.); приводится общая частота словоформы. 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. 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*.[2] В словарь включена ядерная часть этого списка, насчитывающая 3 000 наиболее частотных единиц.

По традиции, сложившейся для изданий такого рода, на страницах словаря представлена рубрика «Интересные факты»: публикуются списки самых популярных слов различных лексических групп (дни недели, погодные явления, цвета, глаголы движения и т.д.), а также самые длинные словоформы и частотный список знаков пунктуации.

[2] Особо отметим, что прилагательные типа *Христов, Петин, Костромской/костромской* относятся к общей лексике.

6429	костюм	2288	плащ
4890	сапог	2179	юбка
3696	пальто	1904	шинель
3696	рубашка	1894	наряд*
3410	куртка	1822	туфля
3396	шапка	1668	рубаха
3126	ботинок	1633	джинсы
3041	платок	1585	перчатка
2962	пиджак	1522	шуба
2955	брюки	1356	мундир
2840	штаны	1251	фуражка
2686	шляпа	1235	свитер
2617	берет	1134	валенок

Таблица 3. Частотный список обозначений одежды и обуви.

В качестве примера в таблице 3 мы приводим частоты имен существительных, обозначающих одежду и обувь. Как можно ожидать, список отражает, с одной стороны, «типичность» элементов гардероба (*валенки* занимают только 26 место в списке), а с другой стороны, их «значимость» при описании внешности человека в текстах (*костюм* – более перцептивно выделенная вещь, чем *ботинки*).

4. Подготовка словарного материала

Базовые списки частотного словаря были получены в автоматическом режиме, при этом использовалась метатекстовая и лексико-грамматическая разметка корпуса. На основе метатекстовой информации были построены и сравнивались между собой частотные списки на отдельных выборках корпуса (по функциональным стилям, по времени создания текста). 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.[3] Это дало возможность собрать данные о частотности не только отдельных словоформ, но и лексем, а также об употребительности тех или иных грамматических категорий. При создании настоящего словаря был использован вариант лексико-грамматической разметки корпуса с автоматическим разрешением морфологической омонимии.

Русский язык как язык с богатым словоизменением создает дополнительные трудности для составителей частотного словаря, так как многие словоформы в текстах омонимичны (ср. словоформу *стали* как форму глагола *стать* и существительного *сталь*, словоформу *банка*, представляющую леммы *банк* и *банка*, слова типа *вера* и *Вера*). Тем не менее, в частотном словаре исходная форма слова, или лемма, 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).

Существенную проблему для лемматизации представляют также несловарные слова (Ляшевская и др. 2007). Если слово отсутствует в грамматической словаре морфологического парсера, то ему приписываются одна или несколько гипотез об исходной форме слова и его грамматических характеристиках. В результате в частот-

[3] Принципы лемматизации и состав частей речи определяются морфологическим стандартом корпуса (НКРЯ 2005), который в общем и целом соответствует принципам Грамматического словаря русского языка (Зализняк 1977). Некоторые особенности лемматизации связаны с тем, что сбор данных происходит по преимуществу в автоматическом режиме. Отметим, что 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.

ный словарь попадают такие «леммы», как *благодарностий* (ср. словоформу *благодарностию*), *Янсный* (ср. *Янсен*), *Барклаивать* (ср. *Барклай*). Между тем, доля несловарных словоформ в НКРЯ составляет 3% всех словоупотреблений и 45% списка словоформ корпуса. Для частотных несловарных словоформ использовались программы пост-обработки морфологической разметки НКРЯ, составленные Б.П. Кобрицовым и Г.К. Бронниковым, а также результаты валидации работы этих программ, полученные О.Н. Ляшевской и Д.К. Бронниковой (Ляшевская 2007, Бронникова 2007). Наиболее эффективными оказались два подхода к лемматизации несловарных слов: кластеризация гипотез о лемме и типе парадигмы (наиболее вероятным для словоформы считается тот разбор, который встречается и у других несловарных словоформ, таким образом, словоформы «ищут» себе соседей по словоизменительной парадигме) и выделение наиболее продуктивных приставок.

Поскольку автоматическое разрешение омонимии и интерпретация несловарных форм допускают определенную, хотя и незначительную, погрешность, омонимы, входящие в первые 20 тысяч частотных слов, подверглись дополнительной ручной проверке.

***

Авторы выражают благодарность В.А. Плунгяну, А.Я. Шайкевичу, а также Е.А. Гришиной, Б.П. Кобрицову, Е.В. Рахилиной, Д.В. Сичинаве и другим участникам семинара НКРЯ, принимавшим участие в обсуждении принципов создания словаря. Мы благодарим О. Урюпину, Д. и Г. Бронниковых, Б. Кобрицова, сотрудников ООО «Яндекс» А. Аброскина, Н. Григорьева, А. Сокирко за помощь в сборе и обработке материала.

Список литературы

1. Бронникова Д.К. Сравнение алгоритмов лемматизации на материале Национального корпуса русского языка. Дипломная работа. М.: РГГУ, 2007.

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

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

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

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

6. Лённгрен Л. (ред.). Частотный словарь современного русского языка [Lönngren, Lennart. The Frequency Dictionary of Modern Russian. Acta Univ. Ups., Studia Slavica Upsaliensia Uppsala 32]. Uppsala, 1993.

7. Ляшевская О.Н.. К проблеме лемматизации несловарных слов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». М, 2007.

8. Ляшевская О.Н., Кобрицов Б.П., Сичинава Д.В. Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика 2007. Екатеринбург, 2007.

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

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

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

12. Сегалович И., Маслов М.. 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.

13. Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Международная конференция «Корпусная лингвистика 2004». С.-Пб., 2004.

14. Степанова Е.М. Частотный словарь общенаучной лексики. М., 1976.

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

16. Штейнфельд Э.А. Частотный словарь современного русского литературного языка. Таллин, 1963.

17. Čermák F., Křen M. (eds.). Frekvenční slovník češtiny (Frequency dictionary of Czech). Praha: NLN, 2004.

18. Čermák F., Křen M. New generation corpus-based frequency dictionaries: The case of Czech // International Journal of Corpus Linguistics, 10, 2005. P. 453-467.

19. Church K.W. 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.

20. Cieri Ch., Liberman M. Language resources creation and distribution at the Linguistic Data Consortium // Proceedings of LREC 02. Las Palmas, Spain, 2002. C. 1327-1333.

21. Davies M. A Frequency Dictionary of Spanish: Core Vocabulary for Learners. London – N.Y.: Routledge, 2005.

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

23. Juilland A., Brodin D., Davidovitch C. Frequency Dictionary of French Words. The Hague-Paris: Mouton, 1970.

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

25. Leech G., Rayson P., Wilson A. Word Frequencies in Written and Spoken English: based on the British National Corpus. London: Longman, 2001.

26. Rayson P., Garside R. Comparing corpora using frequency profiling // Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000. P. 1-6.

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