ФИЛОЛОГИЯ

УДК 371.214: 811.161.1

ЧАСТОТНОЕ РАСПРЕДЕЛЕНИЕ СЛОВ ПО СЕМЕСТРАМ УЧЕБНОЙ ПРОГРАММЫ КАФЕДРЫ РУССКОГО ЯЗЫКА ГГУ

А. К. Голандам
Гилянский государственный университет, г. Решт, Иран

FREQUENCY DISTRIBUTION WORDS FOR SEMESTER CURRICULUM DEPARTMENT OF RUSSIAN LANGUAGE UNIVERSITY OF GUILAN

Arash Golandam
University of Guilan, Resht, Iran

Summary. The article deals with the theory of a frequency distribution of words by semester curriculum Russian Language Department University of Gilan (Iran). The description of the work, as well as the scope of its use.

Key words: frequency dictionary; language level.

Изучение иностранных языков чрезвычайно трудоемкая задача, и разумная методика может сэкономить годы, тогда как плохая делает цель недостижимой. Чем чаще элемент встречается, тем нужнее его знание. Поэтому слова, фразеология, грамматические правила должны изучаться в порядке уменьшения частоты их употребления в тексте (речи). Загромождение учебников случайной лексикой, подчеркнутое выпячивание редких конструкций замедляют изучение иностранного языка.

Частотный словарь включает в себя те слова или другие лингвистические единицы (словоформы, словосочетания), которые зарегистрированы составителем в обследованных им текстах (или тексте). При этих словах, словоформах и т. д. указываются частоты их употребления в данных текстах (тексте) [1, с. 5].

Работа по составлению частотного словаря требует, прежде всего, терпения, усидчивости и определенной лингвистической квалификации. Можно ограничиться чисто формальным определением текстового слова, словоупотребления как цепочки букв от пробела до пробела (как это принято обычно в статистической лексикографии). Соответственно, слово (точнее, словоформа) словаря данного текста будет внешне полностью совпадать со словоупотреблением. Но слово (словоформа) – это одна из разных единиц текста, т. е. единица его словаря, а словоупотребление – это единица самого текста, одна из всех его единиц. Здесь пришлось подробно объяснять привычные для лексикографа-частотника термины, которые нередко смущают «обычного» филолога. Если ограничиться только внешним обликом учитываемой в тексте единицы частотного словаря, то проблем как будто нет никаких. Так что сделать хоть какой-нибудь частотный словарь не так уж трудно. Труднее сделать хороший частотный словарь. Остается лишь решить, какой именно текст (тексты) подвергнуть анализу [1, с. 12]. Понятие «частотный словарь», таким образом, охватывает не столько систему размещения словарного материала, сколько принцип его отбора (в такой словарь попадают только те единицы, которые встретились составителю в проанализированном тексте) и численную регистрацию его фактического употребления [1, с. 15]

Быстрый рост индустрии информационно-поисковых систем, стимулированный расширением сферы Интернета в последние годы, происходит в

условиях слабой развитости автоматизированных средств анализа естественно-языковой информации. Основная проблема связана с недостаточной проработанностью лингвистического обеспечения, однако есть все основания предполагать, что ситуация принципиально не изменится в ближайшее десятилетие как для русского, так и для прочих языков. К настоящему моменту прикладная лингвистика способна обеспечить разработчиков интеллектуальных систем не более чем средствами морфологического и лексического анализа.

Частотный словарь является источником информации о том, какие слова более употребительны в языке, а какие менее частотные. 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. Собранная в частотном словаре информация необходима для многих целей: обучение языку, лингвистические исследования, выбор словников для создания словарей, компьютерные приложения. В нашей работе используется частота слов для целей обучения иностранному языку.

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 множества текстов из учебной литературы, функционирующих в современном русском языке. Данный частотный словарь подобен некоторым частотным словарям (например, Davies, 2005), которые были созданы для преподавания языка. В них не только отражена частотность отдельных слов, но и приводится дополнительная грамматическая и лексическая информация. Целью создания данного словаря было предложить пользователям достаточно представительный базовый словник современного русского языка как иностранного, который можно использовать и адаптировать для разнообразных учебных целей. Данный словарь основан на выборке текстов из учебников РКИ (русский язык как иностранный) и учебной литературы русского языка филологического профиля, представляющей современный русский язык периода 1990–2007 годов. Объем выборки составляет около 15 млн словоупотреблений.

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 придерживается стандарт по составлению частотных словарей. При составлении частотника были приняты и использованы принципы НКРЯ (Национальный корпус русского языка). В ЧРС (частотное распределение слов) тексты классифицируются по нескольким параметрам [2]. Частотный корпус слов получен путем анализа учебников и учебной литературы по русскому языку как иностранному. Были проанализированы учебники РКИ (русский язык как иностранный) по частотному принципу, и отобраны в базу данных слова по частотному убыванию. Полученные слова были распределены по языковому уровню в группы слов, начиная с более частотных и заканчивая менее частотными. Все полученные слова после частотного распределения были разделены на 8 учебных семестров по учебной программе вуза. Нужно было определить

границы каждого уровня; естественно, каждый высший уровень будет содержать слова уровней, находящихся под ним. Таким образом, схема будет такой: 1 уровень = 1000 слов, 2 уровень = первые 1000 слов 1-го уровня + 1000 слов и т. д. Для этого надо выявить примерную границу слов между различными уровнями. Чтобы автор, который работает над книгой, знал, сколько именно слов относятся к определенному уровню и какие слова имеют общую значимость (относятся к другим уровням) + вместе взятым уровням. Для этого после окончания анализа и работы с целью достижения вышеуказанного были сопоставлены найденные данные с Национальным корпусом русского языка. Тексты художественной литературы включают в себя в основном прозу, биографию писателей и поэтов, а также тексты по литературоведению. Нехудожественные тексты делятся на группы по сфере применения: бытовая, официально-деловая, политическая, публицистика, устные тексты, СМИ, учебно-грамматическая, научно-техническая и исламо-философская. Тематика текстов кодируется списком из 35 категорий, имеющих разную степень подробности: от «экономики» или «политики» до «грамматики», «спорта» или «рекламы».

Таблица 1

Функциональные стили подкорпуса

Художественная литература	10 %
Культурно-спортивные тексты	10 %
Учебно-научные	10 %
Официально-деловые тексты	10 %
Исламо-философские	10 %
Научно-грамматические тек-	10 %
СМИ	10 %
Политико-социологические тексты	10 %
Научно-технические тексты	10 %
Устная литература	10 %

Таблица 2

Первые сто частотных слов каждого уровня

1 уровень	2 уровень	3 уровень	4 уровень	5 уровень	6 уровень	7 и 8 уровни
это	что	что	что	что	что	что
что	это	это	это	как	как	как
как	как	как	как	это	это	это
она	она	она	упражнение	его	его	или
когда	когда	когда	она	она	она	его
они	где	где	его	когда	когда	для
нет	студенты	русский	когда	упражнение	они	когда

где	русский	нет	предложения	они	или	она
студенты	нет	студенты	где	мне	упражнение	они
мой	мне	мне	мне	или	для	если
русский	они	они	они	очень	русский	русский
есть	есть	очень	или	где	мне	при
вчера	очень	его	очень	нет	студенты	упражнение
очень	его	есть	студенты	предложения	очень	которые
мне	или	или	нет	студенты	где	так
его	меня	меня	русский	для	нет	где
дома	мой	мой	прочитайте	был	так	все
друг	вам	вам	для	меня	предложения	очень
или	кто	которые	меня	уже	был	мне
кто	вас	кто	есть	было	все	студенты
меня	почему	почему	вам	так	было	нет
сейчас	друг	слова	был	прочитайте	уже	только
вам	которые	вас	слова	русский	которые	было
был	слова	был	которые	задание	меня	был
которые	был	друг	можно	все	есть	может
сегодня	сейчас	для	было	есть	только	уже
читайте	дома	говорит	мой	которые	прочитайте	предложения
говорит	вчера	дома	почему	глаголы	задание	можно
слова	язык	сейчас	так	время	время	есть
слушайте	для	язык	все	слова	можно	время
этот	пожалуйста	можно	вас	кто	кто	быть
потому	потому	вчера	кто	только	человек	меня
вас	можно	человек	человек	этот	глаголы	либо
сестра	сколько	пожалуйста	время	можно	слова	прочитайте
театр	говорит	так	только	мой	этот	кто

преподаватель	этот	потому	говорит	значения	при	человек
язык	так	будет	дома	вчера	почему	этот
куда	диалог	этот	сейчас	вам	вам	задание
текст	текст	сколько	много	этой	значения	слова
будет	какой	если	значения	сейчас	этой	этой
сколько	куда	предложения	чтобы	говорит	говорит	значения
студент	скажите	текст	уже	либо	сейчас	чтобы
знаете	сегодня	университете	друг	надо	вчера	глаголы
брат	предложения	какой	этот	дома	мой	статья
надо	будет	было	язык	куда	были	чем
какой	читайте	сегодня	этой	почему	если	России
были	было	куда	вопросы	кого	чтобы	после
смотреть	том	нужно	глаголы	вас	либо	будет
хорошо	любит	день	будет	чтобы	надо	были
Анна	нужно	диалог	если	язык	дома	почему
моя	здесь	скажите	при	были	России	этом
прочитайте	знаете	чтобы	вчера	при	вас	вам
было	человек	время	либо	человек	язык	язык
автобус	если	читайте	какой	если	кого	том
дом	прочитайте	том	надо	много	куда	сейчас
есть	вопросы	любит	сегодня	друг	чем	говорит
здесь	хорошо	прочитайте	ему	ему	после	вчера
человек	вот	только	может	после	может	мой
там	театр	лет	были	раз	много	надо
письмо	чтобы	надо	сколько	будет	раз	дома
комнате	упражнение	вопросы	нужно	может	друг	ему
партнер	может	хорошо	здесь	день	ему	раз
том	день	уже	хорошо	сегодня	будет	вас
вопросы	кого	какие	раз	чем	день	кого
жив	время	может	потому	том	том	стать

почему	надо	здесь	день	нужно	этом	много
вечером	лет	знаете	скажите	времени	лет	лет
лет	какие	театр	пожалуйста	этом	сегодня	куда
ваш	уже	упражнение	этом	лет	была	друг
хороший	преподаватель	все	лет	сколько	времени	день
знаю	только	вот	какие	университете	нужно	них
читать	слушайте	много	данные	хорошо	всегда	времени
какие	узнать	магазин	внимание	была	сколько	была
Москве	глаголы	были	после	какой	свой	без
друзья	были	кого	текст	нас	какой	свой
любит	все	глаголы	том	всегда	быть	сегодня
вот	Москве	узнать	вот	текст	хорошо	договора
задание	сестра	задание	сказал	свой	нас	нужно
образец	там	сказал	глагол	театр	текст	года
магазин	работать	работать	существительные	образец	вот	какой
работать	много	слушайте	обратите	вопросы	них	всегда
словарь	моя	быть	них	вот	часто	себя
день	смотреть	Москве	университете	здесь	университете	сколько
предложения	сказал	преподаватель	задание	сказал	себя	эти
наш	Анна	сестра	куда	через	театр	через
так	ваш	ему	театр	потому	здесь	жизни
диалоги	быть	квартира	него	него	вопросы	вот
этой	ему	смотреть	работать	быть	через	хорошо
сказал	друзья	там	читайте	глагол	него	часто
него	дом	моя	кого	часто	потому	него
глаголы	брат	музей	была	действия	образец	здесь
пожалуйста	существительные	ваш	всегда	пожалуйста	ситуациях	также
купил	задание	чем	быть	ситуациях	эти	нас

этом	знаю	друзья	работу	городе	жизни	действия
если	большой	больше	знаете	работу	сказал	ситуациях
интересный	интересно	этой	чем	работать	без	текст
часто	жив	спросил	ответьте	стать	какие	работу
большой	раз	дом	свой	них	работать	университете
была	магазин	раз	любит	себя	действия	один
газеты	хороший	другу	словосочетания	эти	пожалуйста	потому

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. 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.

Учитывается только пословная разметка: устойчивые обороты, составные предлоги и другие неоднословные лексические единицы (*Новый год, в течение, тем не менее, друг друга)* не включаются в словарь. Части сложных слов, записанные через дефис (*город-спутник, член-корреспондент),* как правило, учитываются по отдельности. Формы причастий входят в парадигму глагола. Возвратные и невозвратные глаголы, глаголы совершенного и несовершенного вида считаются отдельными единицами словаря.

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, падеж – падёж).* 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. 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.*

Структура словаря

Словарь состоит из следующих разделов:

I. Общая лексика:

- алфавитный список лемм;

- частотный список лемм;
- распределение лемм по функциональным стилям:

– частотные слова художественной литературы;
– частотные слова устной речи;
– частотные слова учебно-научных текстов;
– частотные слова официально-деловых текстов;
– частотные слова исламо-философских текстов;
– частотные слова научно-технических текстов;
– частотные слова научно-грамматических текстов.

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. Имена собственные и аббревиатуры:

- алфавитный список 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 10 000 самых частотных лемм. Частотные словари функциональных стилей составлены на основе подкорпусов художественной литературы, публицистики, научно-грамматических текстов, 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 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.

Библиографический список

1. Алексеев П. М. Частотные словари : учеб. пособие. – СПб. : Изд-во Санкт-Петербургского ун-та, 2001.
2. Автоматическая обработка текста. URL: www.aot.ru.
3. Баранов А. Н. Введение в прикладную лингвистику. – М., 2003.
4. (ред.). Частотный словарь русского языка / под ред. Л. Н. Засориной. – Л. : Наука,1977.
5. Карпова Г. Д., Пирогова Ю. К., Кобзарева Т. Ю., Микаэлян Е. В. Компьютерный синтаксический анализ: описание моделей и направлений разработок // Итоги науки и техники. Сер. «Вычислительные науки». Т. 6. – М., 1991.
6. Корпус Русского Языка 2003–2005. Результаты и перспективы. – М. : Индрик. – С. 6–20.

7. Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой; Ин-т языкознания АН СССР. – М.: Сов. энцикл., 1990.
8. Маслов Ю. С. Введение в языкознание. – Л., 1987.
9. НКРЯ: Национальный корпус русского языка 2003–2005: Результаты и перспективы. – М.: Индрик, 2005.
10. Савчук С. О. (2005). Метатекстовая разметка в Национальном корпусе русского языка // Национальный Корпус Русского Языка 2003–2005. Результаты и перспективы. – М. : Индрик, 2005. – С. 62–88.
11. Статистика речи и автоматический анализ текста / отв. ред. Р. Г. Пиотровский. – Л., 1980.
12. Степанова Е. М. Частотный словарь общенаучной лексики. – М., 1976.
13. Шаров С. А. Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Сер. 2. – 2003. – № 6. – С. 9–17.
14. Шемакин Ю. И. Начала компьютерной лингвистики : учеб. пособие. – М., 1992.
15. Штейнфельд Э. А. Частотный словарь современного русского литературного языка. – Таллин, 1963.