А. Я. ШАЙКЕВИЧ,
В. М. АНДРЮЩЕНКО, Н. А. РЕБЕЦКАЯ

СТАТИСТИЧЕСКИЙ СЛОВАРЬ ЯЗЫКА РУССКОЙ ГАЗЕТЫ (1990-е годы)

Том 1

STUDIA PHILOLOGICA

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ РУССКОГО ЯЗЫКА им. В. В. ВИНОГРАДОВА

А. Я. ШАЙКЕВИЧ,
В. М. АНДРЮЩЕНКО, Н. А. РЕБЕЦКАЯ

СТАТИСТИЧЕСКИЙ СЛОВАРЬ ЯЗЫКА РУССКОЙ ГАЗЕТЫ (1990-е годы)

Том 1

ЯЗЫКИ СЛАВЯНСКИХ КУЛЬТУР
Москва 2008

ŋqdqɏeɎ bɆ dedɆbɎɜə d ɕɔɑbɔɏ Ɉqddɏqɏè

ББК 83
Ш 12

Издание осуществлено при финансовой поддержке
Российского фонда фундаментальных исследований
(РФФИ)
проект № 08-06-07023

Рецензенты:
доктор филол. наук *Д. О. Добровольский*,
доктор филол. наук *С. Е. Никитина*

А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая

Ш 12 Статистический словарь языка русской газеты (1990-е годы) / Рос. акад. наук. Ин-т русского языка им. В. В. Виноградова. Т. 1. — М.: Языки славянских культур, 2008. — 592 с., разд. паг. (VIII, 578 с.). — (Studia philologica).

ISSN 1726-135X
ISBN 978-5-9551-0279-5

Настоящий словарь представляет собой композицию трех частотных словарей, за каждым из которых стоит свой корпус текстов: 1) тексты девяти московских и петербургских газет за 1997 г., 2) комплект «Независимой газеты» за 1996—2000 гг., 3) корпус газетных заголовков 1991—2000 гг. Общий объем трех корпусов составляет 50 млн слов текста. В печатной версии словаря представлено 52 тыс. разных слов, в электронной версии словник превышает 140 тыс. разных слов, там же читатель найдет и соответствующий обратный словарь.

Во втором томе словаря будут даны таблицы распределения лексики по газетам, жанрам и темам; электронная версия включит таблицы бинарных словосочетаний.

ББК 83

ISBN 978-5-9551-0279-5

Введение

Жанр частотных словарей когда-то рассматривался лингвистами как сугубо прикладное направление. Один из грандиознейших словарей подобного рода, созданных в докомпьютерную эру, вышел в свет под характерным заголовком — «Словарь учителя...» (Thorndike E. L., Lorge I. The Teacher's Word Book of 30,000 words. N. Y., 1944). Первый частотный словарь русского языка (Josselson H. The Russian Word Count. Detroit, 1953) был адресован преподавателям, он опирался на корпус в 1 млн словоупотреблений (1830–1950) и содержал более 5 тысяч разных лексем. На основе корпуса в 400 тыс. словоупотреблений (литература для детей и юношества) построен «Частотный словарь современного русского языка» Э. А. Штейнфельдт (Таллин, 1963), в нем было представлено 2500 наиболее употребительных слов.

Первый опыт использования компьютера для этих целей материализовался в виде «Частотного словаря русского языка» под ред. Л. Н. Засориной (М., 1977). Словарь построен на базе текстов 1900–1960-х годов общим объемом 1 млн словоупотреблений. Опубликованный словарь включал весь лексический материал (кроме имен собственных). В 1993 г. в Швеции опубликован «Частотный словарь современного русского языка» Л. Леннгрена (Uppsala, 1993). Исходный корпус (1 млн словоупотреблений) содержал в равной доле художественную литературу 1960–1980-х годов и журнально-газетные тексты 1985–1988 годов. К сожалению, в публикацию не включены лексические единицы с частотой 1–9, т. о. в печатном словаре находим около 9 тыс. лексем.

Появление персонального компьтера, сканирующих устройств, развитие Интернета было технологической революцией и, казалось бы, обещало появление все новых и новых частотных словарей. В действительности, прогресс в этой области был не столь быстрым. При росте текстовых корпусов на два (или даже три порядка) автоматизация разметки текста и лемматизации повысли эффективность труда на порядок.

Первым примером такого типа может служить Частотный словарь чешского языка — Frekvenční slovník češtiny (pod vedením F. Čermáka a M. Křena). Praha, 2004. Национальный корпус чешского языка, на базе которого создан этот словарь, включает около 100 млн. словоупотреблений. В словаре представлено более 50 тыс. лемм, начиная с частоты 13.

Одновременно меняются конечные цели, ради которых создаются статистические словари языка. Во главу угла ставится уже не прикладная задача отбора лексики для студентов-иностранцев, но фундаментальная задача описания языка во всем разнообразии жанров, стилей и периодов развития. Первым замечательным образцом такого рода был «Dictionnaire des fréquences (P.: Didier, 1971). Соответствующий корпус в 70 млн словоупотреблений включал тексты художественной литературы от 1789 до 1964 г. В итоговых таблицах отражены 4 хронологических среза.

Настоящий словарь должен положить начало серии статистических словарей для разных периодов и разных жанров русского языка. Замысел возник еще в начале 1990-х гг. В те «тощие» для науки годы в Отделе машинного фонда Института русского языка им. В. В. Виноградова РАН удалось вручную ввести в компьютер заголовки нескольких газет 1991–93 гг. («Вечерняя Москва», «Известия», «Комсомольская правда», «Московская правда», «Независимая газета», «Правда», «Сегодня»). Позднее к ним были добавлены заголовки электронных корпусов, упоминаемых ниже. Общий объем корпуса заголовков составляет 1,5 млн словоупотреблений. В первом томе нашего словаря соответствующие частоты даны в правом столбце.

К началу 1997 г. ситуация начала меняться. Появилась возможность подписки на электронные издания газет, что, впрочем, потребовало продолжительной неустанной борьбы с опечатками. **Сама эта работа стала возможной благодаря финансовой поддержке РФФИ (гранты 97–06-80100, 00-06-80230, 04-06-80094).** Так возник корпус девяти газет 1997 года (в основном, второй половины года): «Известия" (2146 тыс. словоупотреблений), «Литературная газета» (1043), «Московский комсомолец» (1887), «Независимая газета» (3536), «Новая газета» (419), «Правда-5» (737), «Российские вести» (706), «Санкт-петербургские ведомости» (1614), «Сегодня» (2423). В этот корпус были добавлены тексты из националистических изданий 1991–93 гг. («Завтра» и т. п. — 231), из «Московских новостей» 1995 г. (201) и «Литературной газеты» 1994 г. (45). общий объем результирующего корпуса составил 14987 тыс. или круглым счетом 15 млн словоупотреблений. В настоящем словаре этот корпус условно именуется «Корпусом 1997», соответствующие частоты приводятся в левом столбце.

Компакт-диски с электронными комплектами «Независимой газеты» стали появляться с 1998 г. Стало возможным появление третьего газетного корпуса — текстов «Независимой газеты» за 1996–2000 гг.общим объемом 35 млн словоупотреблений. Соответствующие данные образуют средний столбец наших таблиц. На материале этого последнего корпуса можно впервые начать анализ краткосрочной лексической динамики, что методически важно для составителей частотных словарей, кроме того, материал этот может заинтересовать историков и политологов.

Таким образом, публикуемый словарь представляет собой композицию трех частично пересекающихся частотных словарей. В левый столбец включено 10% материала среднего столбца (вторая половина 1997 г.), большая часть материала правого столбца присутствует в двух других столбцах. Суммирование данных трех столбцов, следовательно, неправомерно.

Словник

В нашем материале представлено 811 тысяч разных графических слов. Чтобы сделать осмысленным это богатство форм, необходимо свести его к меньшему числу вокабул. В громадном большинстве случаев (но не во всех) вокабула совпадает с леммой, т. е. со словарной формой представления слова. О принципах лемматизации речь пойдет ниже, пока же остановимся на проблеме ограничения объема словника. Для печатной версии словаря вводим следующие правила:

1. Слова русского языка (за исключением имен собственных), их аффиксальные производные и сложные слова со слитным написанием включаются в словарь, начиная с частоты 13.

2. а) Энклитические частицы с дефисным написанием (*-де, -ка, -кась, -от, -с, -таки, -тко, -то*) отделяются от предшествующего слова и включаются в словарь без частотных ограничений. Также разделяются слова с конечным *-другой*.

б) Разделение слова с дефисом не проводится, если частица *-то* присоединяется к местоименному слову с начальным *к-* (*г-*) или *ч-*.

в) Сложные числительные с дефисным написанием разделяются на два слова. Также разделяются дефисные написания с названиями месяцев.

г) Не разделяются на части графические слова *вообще-то, все-таки, наконец-то, ну-ка, ну-с, ну-тка, опять-таки, потому-то, столько-то, сякой-то, так-то, такой-то, то-то, туда-то, тут-то*.

3. Имена собственные включаются в словарь, если их частота превышает 29. Этот же порог действителен для аббревиатур организаций, компаний и фирм, всевозможных мероприятий с цифровым компонентом (*Ту-130, Формула-1* и т. п.).

4. Леммы, записанные латиницей, включаются в словарь, начиная от частоты 5.

5. В словарь включены и числа, если их частота превышает 99.

Лемматизация

Строгое разделение частей речи потребовало бы изощренной программной системы или ручного обследования миллионов контекстов. Особенно трудоемко разделение прилагательных и причастий, в несколько меньшей мере прилагательных и существительных. Далеко не все эти задачи имеют однозначное решение.

При составлении данного словаря мы следовали гибкой политике, иногда жертвуя грамматической чистотой в пользу лексической семантики. Синтетические компаративы и суперлативы выделялись в особые вокабулы, равно соотнесенные с прилагательными и наречиями. Деадъективные формы на *-о* считались особыми вокабулами (в них может быть скрыт какой-то процент кратких форм среднего рода прилагательных).

Стремление к посильной семантической дифференциации заставляло иногда расщеплять традиционные словарные статьи обычных словарей. Так, разделены формы числа существительных *Бог* и *боги*, *выбор* и *выборы*, *круг* и *круги*, *многое* и *многие*, *новости* и *новость*, *рамка* и *рамки*, *сведение* и *сведения*, *цвет*, *цветок* и *цветы*. Трактуется как особая вокабула форма *образом*. Отдельными вокабулами стали *болит* (*болело*), *давай*(*те*), *кажется*, *придется* (*пришлось*), *приходится* (*приходилось*), *разумеется*, *следует* (*следовало*), *сообщается*, *спрашивается*, *стоит*

(*стоило* — как модальное слово), *хватает* (*хватало*) и *хватит* (*хватило*). При таком расщеплении учитывлись как семантические, так и частотные соображения. Большинство выделенных глагольных вокабул встречаются намного чаще суммы всех остальных форм соответствующего глагола.

Особыми вокабулами являются многие краткие формы прилагательных: *велик, намерен, нужен, обязан, принято, равен, согласен, таков, убежден*. Довольно часто в отдельные вокабулы выделялись потенциальные причастия (частично адъективированные или субстантивированные): *действующий, заведующий, интригующий, командующий, кричащий, курящий, любящий, манящий, мертвящий, моющий, мыслящий, мятущийся, обжигающий, облитерирующий, ободряющий, объемлющий, окружающий, освежающий, отдыхающий, отравляющий, отягчающий, ошеломляющий, палящий, подавляющий, подобающий, подрастающий, подходящий, порочащий, последующий, потрясающий, правительствующий, правящий, практикующий, предержащий, предстоящий, предшествующий, предыдущий, преходящий, привходящий, прилегающий, прилежащий, приличествующий, проезжающий, проникающий, присутствующие, профилирующий, процветающий, пьянящий, решающий, руководящий, соответствующий, страждущий, странствующий, текущий, трудящийся, тяжущийся, угрожающий, удручающий, ужасающий, укоряющий, упреждающий, успевающий, устрашающий, учащийся, фашиствующий, хрустящий, хулиганствующий, цветущий, чарующий; минувший, прошедший; обвиняемый, предполагаемый; направленный, обеспеченный, обреченный, обусловленный, основанный, построенный, предназначенный, представленный, предусмотренный, расположенный, распространенный, связанный, сокращенный, указанный.*

В какой-то мере разделение на вокабулы проводилось в рамках одной части речи исключительно по семантическим соображениям. Таковы *лев, лев* (болг.) и *Лев*; *среда* и *среда* (день недели), *статья* и *статья* (журн.).

Словосочетания

Отличительной чертой настоящего словаря можно считать включение в него в качестве вокабул не только отдельных слов, но и словосочетаний. Нам известен только один частотный словарь, дающий сочетания слов (collocations). Это Allen S. e. a. Frequency Dictionary of Present-day Swedish based on newspaper material. Uppsala, 1970–75. Статистике словосочетаний посвящен 3-й том словаря, где приводятся все бинарные словосочетания с частотой 2 и более (включая самые тривиальные грамматические сочетания). Словарь Аллена построен на корпусе 1 млн словоупотреблений.

В таблицу 1 настоящего словаря включено несколько десятков словосочетаний, таких, как: *в качестве, в конечном счете, в конце концов, в связи с, в противном случае, в случае, в соответствии с, в состоянии, в сущности, в течение, в том числе, в угоду, в ходе, в частности, вверх тормашками, во избежание, вряд ли, до свидания, до сих пор, друг друга, за неимением, за счет, и так далее, и тому подобное, исполняющий обязанности, к сожалению, круглый стол, между тем, может быть, на плаву, на скаку, на протяжении, населенный пункт, настоящее время, по крайней мере, по поводу, по сравнению с, по существу, повестка дня, потому что, права человека, с поличным, сломя голову, стало быть, судя по, Счетная палата, табель о рангах, так как, так называемый, так что, таким образом, тем не менее, тем самым, то есть, только что, тот самый, точка зрения, Филькина грамота, что касается.*

Во втором томе настоящего словаря будут представлены все бинарные словосочетания с частотой 3 и более. Приведем несколько (лемматизированных) примеров с левым прилагательным. Числа указывают на частоту словосочетания, числа в скобках - частота прилагательного.

кричащий [82] заголовок 4, проблема 3, противоречие 3
моющий [21] средства 11
мятущийся [46] душа 9
мыслящий [292] люди 27, политик 9, человек 9, лидер 5, часть 5 интеллигенция 3
отравляющий [88] вещество 50, газ 5
ошеломляющий [87] успех 15 впечатление 8 победа 3

Обратимся к более частым словам:

ђqфqbɐҮ qƎ dɐdƎqҮɜɐ d ɐɔɐqɔҮ ɹqфqҮqҮè

действующий [5766]:
ныне 30, сейчас 15, сегодня 11, уже 5;
законодательство 457, конституция 380, закон 178, соглашение 18, устав 11, документ 10, договор 9, УК 9, программа 8, контракт 5, редакция 5;
система 54, модель 22, механизм 21, норма 20, порядок 18, режим 11, ставка 11, правило 10, схема 8, бюджет 8, санкции 4, тариф 4;
президент 400, губернатор 231, председатель 174, глава 107, мэр 35, премьер 18, руководитель 13, премьер-министр 9, лидер 9, министр 9, градоначальник 5;
власть 96, политик 49, правительство 37, депутат 30, парламент 22, кабинет 14, руководство 7, состав 6;
орган 35, структура 13, круглый стол 5;
сотрудник 15, член 12, генерал 7, офицер 6, чиновник 5;
чемпион 10;
храм 11, церковь 4;
предприятие 18, газопровод 8, шахта 7, АЭС 6, банк 5, месторождение 5, выставка 5, объект 4, станция 4, блок 4, скважина 4;
субъект 7;
фактор 7;
армия 34;
лицо 415.

Приведенные примеры сгруппированы по смыслу и в своей совокупности семантически характеризуют данное прилагательное. Обратим внимание на сравнительно новое словоупотребление — *действующий чемпион*. Мы видим здесь также кандидата во фразеологизмы - *действующая армия* и безусловный фразеологизм - *действующее лицо*.

Прилагательное *действующий* лишено каких-либо эмоциональных коннотаций. Напротив, прилагательное (или причастие) *правящий* ярко окрашено стилистически, и правые компоненты словосочетаний это демонстрируют.

правящий [2504]:
партия 491, коалиция 336, элита 255, режим 123, круги 119, класс 113, верхушка 60, альянс 41, слой 39, большинство 38, группа 20, группировка 18, блок 18, династия 15, бургомистр 12, олигархия 11, архиерей 10, кабинет 10, лагерь 10, ныне 10, номенклатура 9, семья 8, команда 7, бюрократия 3, епископ 6, курс 6, структура 6, власть 5, президент 5, клан 4, монарх 4, истеблишмент 3, пока 3,триумвират 3.

Исследователь-лингвист многое почерпнет из статистики, казалось бы, самых простых сочетаний. Вот два примера предложного управления:

на Украине	2504	в Украине	142
на Украину	618	в Украину	25
на Охотном	213	в Охотном	48

Мы видим, что «Независимая газета» в 1990-х годах держится традиции предложного управления словом *Украина*, но явно отказывается от традиционного московского *в Охотном ряду*.

Хронология

В словарь включена хронологическая таблица текстов «Независимой газеты». Ниже приводится несколько примеров различных тенденций динамики частоты:

	1996	1997	1998	1999	2000
	Монотонное нарастание частоты				
Евросоюз/ЕС	504	676	997	974	1663
компания	2904	3656	4225	4117	4759
холдинг	37	77	229	270	509
СМИ	776	1155	1289	1506	2041
Кремль	1164	1338	1341	1636	2427
чиновник	938	1225	1298	1069	1308
номинация	194	291	314	331	341
суд	2813	2880	3384	3827	4389
(ген)прокуратура	907	986	1190	1370	1957
экстремист/-зм	283	224	480	749	889
фигурант	18	18	36	69	53
Березовский	173	458	725	790	854
	Монотонное падение частоты				
СССР	2976	2160	1821	1712	1874
оппозиция	2736	2440	2108	1723	1668
революция	1052	927	841	763	772
джаз	160	144	43	79	68
	Пик частоты в 1996 г.				
Ельцин	8485	5678	5302	4238	2069
консилиум	51	14	9	17	6
теннис	93	54	58	57	40
Гайдар	573	467	518	313	148
КС	499	497	339	147	192
Лебедь	3360	913	1105	414	212
Зюганов	3165	698	1032	837	763
	Пик частоты в 1997 г.				
секвестр	2	323	53	10	14
Диана	33	164	58	40	38
празднование	152	273	211	200	245
	Пик частоты в 1998 г.				
доллар	3692	3916	5290	4443	4752
рубль	3327	3943	5019	3377	3337
правительство	9458	10810	13916	10652	9455
отставка	1172	1356	1846	1750	1222
	Резкий подъем частоты в 1998 г.				
олигарх	3	27	496	238	842
	Пик частоты в 1999 г.				
МВФ	651	557	1098	1729	781
дефолт			90	230	136
Примаков	788	1115	2497	3637	614
Степашин	79	86	439	2299	297
импичмент	82	94	454	758	70
одномандатный	43	40	108	391	109
Косово	8	8	1164	2812	1068
Пушкин	723	722	773	1488	541

Оглавление