



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

СТАТИСТИЧЕСКИЙ  
СЛОВАРЬ  
ЯЗЫКА ДОСТОЕВСКОГО



STUDIA PHILOLOGICA



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

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

# СТАТИСТИЧЕСКИЙ СЛОВАРЬ ЯЗЫКА ДОСТОЕВСКОГО



ЯЗЫКИ СЛАВЯНСКОЙ КУЛЬТУРЫ  
Москва 2003

**Рецензенты:**

доктор филол. наук *В. П. Григорьев*,  
доктор филол. наук *Н. Н. Перцова*

**А. Я. Шайкевич, В. М. Андрущенко, Н. А. Ребецкая**

Ш 12      Статистический словарь языка Достоевского / Рос. акад. наук. Ин-т русского языка им.  
В. В. Виноградова. – М.: Языки славянской культуры, 2003. – 880 с., разд. паг. (XLVIII, 832 с.). — (Studia  
philologica).

ISSN 1726-135X  
ISBN 5-94457-145-4

«Статистический словарь языка Достоевского» включает всю лексику трех основных жанров писателя — художественной литературы, публицистики и писем (более 43 тысяч разных слов). Словарь построен на корпусе текстов в 2,9 млн. словоупотреблений и значительно превосходит по объему любые другие частотные словари русского языка. По степени лингвистической дифференциации Словарь уникален и в мировом масштабе. В таблицах Словаря лексика Ф. М. Достоевского представлена в распределении по основным жанрам и по периодам творчества. Словарь адресован филологам и всем любителям творчества Ф. М. Достоевского.

**ББК 83**

*Анатолий Янович Шайкевич*

*Владислав Митрофанович Андрущенко*

*Наталья Александровна Ребецкая*

**СТАТИСТИЧЕСКИЙ СЛОВАРЬ ЯЗЫКА ДОСТОЕВСКОГО**

Издатель А. Кошелев

Корректор В. В. Мачкова

Художник-консультант Л. М. Панфилова

Подписано в печать 16.08.2003. Формат 84×108 1/16.

Бумага офсетная № 1, печать офсетная

Усл. печ. л. 89,1. Заказ № 1998. Тираж 550 экз.

Издательство «Языки славянской культуры»

ЛР № 02745 от 04.10.2000.

Тел.: 207-86-93. Факс: (095) 246-20-20 (для аб. М153).

E-mail: lrc-kozlov@mm-net.ru

Каталог в ИНТЕРНЕТ <http://www.lrc-press.ru>; <http://www.lrc-mik.narod.ru>

ISBN 5-94457-145-4



9 785944 571458 >

ГУП Московская типография № 2  
Министерства Российской Федерации  
по делам печати, телерадиовещания  
и средств массовых коммуникаций (МПТР России).  
Тел.: 282-24-91. 129085, Москва, пр. Мира, 105.

© Авторы, 2003

Электронная версия данного издания является собственностью издательства,  
и ее распространение без согласия издательства запрещается.

# Оглавление

Введение	VI-XLVIII
Таблицы	
1. Распределение лексем по основным жанрам	1
2. Относительная частота лексем в основных жанрах	479
3. 100 самых частых лемм в текстах Достоевского	515
4. 100 самых частых лемм в художественных произведениях Достоевского	516
5. 100 самых частых лемм в критике и публицистике Достоевского	517
6. 100 самых частых лемм в письмах Достоевского	518
7. 40 самых частых существительных в основных жанрах	519
8. 40 самых частых глаголов в основных жанрах	520
9. 40 самых частых полных прилагательных в основных жанрах	521
10. Частотный спектр рангового словаря лемм всего корпуса текстов	522
11. Частотный спектр рангового словаря лемм беллетристики	523
12. Частотный спектр рангового словаря лемм критики и публицистики	524
13. Частотный спектр рангового словаря лемм писем	525
14. Лексические маркеры беллетристики	526
15. Лексические маркеры критики и публицистики	537
16. Лексические маркеры писем	550
17. Лексические маркеры беллетристики с максимальными значениями S	561
18. Лексические маркеры критики и публицистики с максимальными значениями S	562
19. Лексические маркеры писем с максимальными значениями S	563
20. Грамматические классы лемм и аффиксы	564
21. Относительная частота грамматических классов лемм и аффиксов	572
22. Распределение лексем в беллетристике по периодам творчества	580
23. Распределение лексем критики и публицистики по периодам творчества	684
24. Распределение лексики писем по периодам жизни Достоевского	757



## Введение

Настоящий Словарь подготовлен в Отделе машинного фонда Института русского языка РАН. Работа над Словарем начиналась в рамках проекта «Словарь языка Достоевского», (руководитель — Ю. Н. Караулов), поддержанного РГНФ, а затем выделилась в самостоятельное направление. На этом этапе авторы использовали также финансовую поддержку РГНФ, оказанную более широкому проекту «Дистрибутивно-статистическое описание языка русской прозы 1855–1880 гг.» (01-04-00247а).

Следует с самого начала подчеркнуть, что цели обоих словарей не совпадают. Цель «Словаря языка Достоевского» — показать лексику писателя во всем ее богатстве (с детальной семантической разработкой, с собранием иллюстративных примеров, с исчерпывающим словоуказателем и т. п.). Итогом является лексикографическая серия, намного превосходящая по объему лучшие образцы авторской лексикографии, такие как первый опыт на русской почве — «Словарь языка Пушкина» [Словарь Пушкина] или замечательный «Словарь языка Мицкевича» [Słownik]. Первый выпуск этой серии уже вышел в свет [Словарь Достоевского].

Задача «Статистического словаря языка Достоевского» скромнее, он должен представить лексику Достоевского в количественном виде, повторив и обогатив опыт уникального конкорданса к Шекспиру [Sprevack]. Однако и при таком ограничении результат оказался бы слишком объемным для бумажного издания (речь идет о многих сотнях авторских листов), а потому было принято решение издать Словарь в гибридном виде — как однотомную книгу, показывающую лишь часть таблиц, и как сопровождающий ее компакт-диск, содержащий информацию в полном объеме. Конечно, в первом опыте такого рода нас подстерегают многие технические трудности издания, а также психологические предубеждения читателей, но именно на этом пути нам видится дальнейший прогресс академической лексикографии.

Предваряя описание структуры Словаря, выскажем одно замечание относительно развития статистической лексикографии. В 1960–1970-х гг. наблюдалось широко распространенное увлечение частотными словарями, особенно в связи с педагогическими и информационными приложениями. От очень скромных по объему изданий (400 тыс. словоупотреблений) лексикография шагнула к рубежу в 1 млн. словоупотреблений, а затем и к новым рекордам — максимально дифференцированный словарь американских текстов для школы содержит более 5 млн. словоупотреблений [Carroll], а словарь, созданный Институтом французского языка [Dictionnaire], опирается на корпус литературных текстов объемом более 70 млн. словоупотреблений. Затем наступает кризис: электронные корпуса текстов продолжают множиться и увеличиваться по объему (в некоторых из них счет идет уже на сотни миллионов словоупотреблений), но не видно новых частотных словарей, которые были бы созданы на основе этих корпусов. В чем же дело? Причин может быть много, назовем некоторые из них.

1) Программными средствами можно легко и просто получить статистику графических слов. Именно такая информация представлена в вышеупомянутом словаре Керрола [Carroll]. Но читателю обычно нужно большее — графические слова должны быть сведены в осмысленные лингвистические единицы, они должны быть лемматизированы. Процесс же лемматизации не поддается алгоритмам на сто процентов. Доля ручного вмешательства хотя и уменьшается относительно, но продолжает расти абсолютно. При росте объема текстового корпуса в 100 раз объем ручного труда при постредактировании возрастет, скажем, в 10 раз.

2) До сих пор не разработаны хорошие автоматизированные процедуры формирования выборки на большом корпусе текстов. Впрочем, эта трудность не

существует при обработке замкнутого корпуса целиком (как, например, в случае текстов Достоевского).

3) Наконец, существует и психологический фактор. Лингвостатистика, как она складывалась в середине XX в., в какой-то степени была во власти математического фетишизма: открытие «закона» Ципфа создавало иллюзию новой области статистических исследований, возникала новая дисциплина, все более терявшая связи с лингвистикой, филологией, информатикой.

Предлагаемый Словарь должен сделать шаг в обратном направлении.

## 1. Корпус текстов Достоевского и его членение

Настоящий Словарь опирается на 30-томное академическое издание Ф. М. Достоевского и в основном следует принципам классификации текстов, принятым в этом издании, т. е. включает три основных жанра — «Художественная литература», «Критика и публицистика» и «Письма». Эти три жанра в совокупности и составляют корпус текстов Достоевского, послуживший базой для всех статистических таблиц «Статистического словаря языка Достоевского». Общий объем корпуса — 2889 тыс. графических слов<sup>1</sup> (145980 разных графических слов), в том числе: «Художественная литература» — 1835 тыс. слов (110744 разных графических слова), «Критика и публицистика» — 524 тыс. слов (59446 разных графических слов), «Письма» — 531 тыс. слов (43689 разных графических слов). Не вошли в наш корпус текстов ранние редакции и варианты, подготовительные материалы и тексты записных книжек. Применение статистических методов к подобным текстам было бы почти невозможным. Не вошли в корпус и деловые бумаги, где индивидуальность автора почти не проявляется. Разумеется, эти группы текстов должны учитываться при составлении исчерпывающего словника Достоевского.

Ряд текстов из «Дневника писателя» отнесен к художественной литературе: «Бобок», «Кроткая», «Мальчик у Христа на елке», «Мужик Марей», «Сон смешного человека», «Столетняя».

## 2. Лингвистические единицы, отраженные в статистических таблицах

В настоящем Словаре представлены как исходные графические слова (только в электронной части Словаря), так и результаты всевозможных процедур над графическими словами (слияние разных грамматических форм слова, слияние вариантов, расщепление, объединение в одну единицу двух и более графических слов, следующих друг за другом). Прежде всего речь идет об орфографических вариантах (*адрес и адресс, прощание и прощанье*), в которых могли проявляться орфографические нормы времени или пристрастия издателей. Подобные варианты объединяются в одну единицу. С другой стороны, сохранена статистическая информация о таких вариантах, как *бриллиант* и *брильянт*, *Авдотья Сергеевна* и *Авдотья Сергевна*, *вести* и *весть*.

Некоторые графические слова разделяются на две или даже три леммы. Речь идет о частицах вроде *-де*, *-ка*, *-с*, *-таки*, *-то*. Однако сохраняются нерасчлененными слова с «неопределенным» *-то*, присоединяемым к основам вопросительных (и некоторых указательных) местоимений (*где-то*, *какой-то*, *откуда-то*, *такой-то*).

Что касается грамматических форм изменяемых слов, то здесь доминирует традиционное представление о частях речи (например, графические слова на *-о*, вроде *абсурдно*, *бездарно*, *безобразно*, *вековечно*, расщепляются на наречия и прилагательные). Однако, вслед за Словарем Пушкина, компаративы сохраняются как отдельные грамматические единицы (при этом формы на *-ее* и *-ей* сливаются воедино). Отдельно фигурируют и суперлативы.

<sup>1</sup> Термин «графическое слово» представляется более правильным, чем общепринятый термин «словоформа». Один раз встретившееся у Достоевского слово *взяточка-то-с* заслуживает названия «графическое слово», но вряд ли будет идентифицировано лингвистами как особая «словоформа». Точно так же графическое слово *ви-но-ват*, встретившееся три раза, едва ли кем-либо будет объявлено особой словоформой. С другой стороны, встретившаяся последовательность *по...за...буду...* («Белые ночи») в словаре графических слов будет отражена как три слова, в словаре лемм — прибавит единицу к частоте слова *позабыть*.



## VIII

Последовательное системное разделение грамматических форм по частям речи в зависимости от синтаксической функции было бы слишком трудоемким. Здесь был принят неформальный принцип: если есть основания предполагать, что синтаксические функции форм отягощены еще и семантическими или стилистическими различиями, они должны быть разведены в статистических таблицах. Сохранена статистическая информация о формах числа многих существительных (*око и очи, ухо и уши, вода и воды, брат и братья* и т. п.), о формах империатива многих глаголов (*не беспокойтесь, ступай* и т. п.). Часто различаются субстантивированные прилагательные (*артельный, блаженный, ближний, больной, большие, бывшее, дворовый, знакомый, рабочий* и т. п.) и омонимичные исходные прилагательные. Выделены и многие адъективированные причастия, например *благодарящий, верующий, воинствующий, волнуемый, заплывший, исхудавший, минувший, обрусевший* и т. п. В максимальной степени грамматическая информация дана для глагола *быть*, для которого указана совокупная частота форм прошедшего времени (представлена формой *был*) и форм будущего времени (представлена формой *будет*).

Выше упомянутый неформальный принцип распространяется и на случаи семантического расщепления, и на случаи объединения последовательностей слов в особые единицы. Так разведены: *а* (союз), *а* (междометие), *а* (вопросительное слово) и *а* (буква); *батюшка* (отец), *батюшка* (обращение), *батюшка* (священник); *благо* (сущ.) и *благо* (союз); *брак* (супружество) и *брак* (дефект) и т. п. Сохранена статистическая информация о словах вроде *акт* (церемония), *банк* (игра), *брат* (обращение), *будет* (достаточно) и т. п.

Довольно часто в статистических таблицах даются сочетания слов, например: *так и быть, была не была, как есть, так и есть, что ни есть, все равно, прежде всего, вещественные доказательства, порядок вещей, в порядке вещей, взад да вперед, взад и вперед, на взгляд, на первый взгляд, по первому взгляду, с первого взгляда, быть молодцу не укор*. Отдельно показаны все имена собственные, в том числе имена с отчествами.

Любую строчку в статистических таблицах настоящего Словаря будем называть **лексемой**. Следовательно, такие строки, как *порядок вещей, в порядке вещей, Аглая, Аглая Ивановна, Адрианополи (город), «Адрианополи» (гостиница), благо (сущ.), благо (союз)*, — все это лексемы. Те лексемы, чьи частоты не входят в частоту других лексем, будем называть **леммами**. Лексемы, не являющиеся леммами, печатаются в таблицах с отступом. Обращаясь, например, к таблице 1, мы найдем там:

	Всего	Х	К	П
Бог	1730	1081	173	476
Бог в помощь	1	1		
Бог ведает	2	1		1
Бог весть	1	1		
Бог знает	322	207	34	81
«Бог»	2		2	
боги	36	25	9	2
ради Бога	459	137	2	320
слава Богу	137	90	11	36

Частоты сочетаний *Бог в помощь, Бог ведает, Бог весть, Бог знает* уже учтены в строке *Бог*, но в ней не учтены частоты лемм *«Бог»*, *боги*, *ради Бога*, *слава Богу*. Если читатель не согласится с такой лемматизацией и захочет получить частоту слова *Бог* в рамках лексикографической традиции, он сможет суммировать частоты этих четырех лемм и получит строку:

Бог	2364	1333	197	834
-----	------	------	-----	-----

Если же, напротив, читатель захочет повысить статус словосочетания *Бог знает*, превратив его в отдельную лемму, ему надо вычесть частоту словосочетания из частоты леммы, получая для леммы *Бог* строку:

Бог	1408	874	139	395
-----	------	-----	-----	-----

Таким образом, различие лексем (вообще) и лемм (в частности) не принципиально – при любом решении статистическая информация сохранена.

### 3. Типы статистических таблиц, представленных в Словаре

Преобладающий тип статистической таблицы (примером может служить таблица 1) содержит текстовую часть, включающую лингвистические объекты: графические слова, лексем (как в таблице 001)<sup>2</sup>, леммы (как в таблице 003), и цифровую часть, состоящую из одного или нескольких столбцов (в таблице 1 – четыре столбца). Как правило, строки лемм в таблице упорядочены по обычному алфавитному принципу.

Исключениями являются обратные частотные словари, в которых единицы упорядочены по алфавиту, как если бы они читались справа налево. Примером может служить таблица 001.

Таблица 001

#### Фрагмент обратного частотного словаря графических слов

52	ба	5	пошиба	7	проба	53	свадьба
2	б-ба	19	лба	1	утроба	2	усадьба
111	баба	6	столба	42	особа	5	ходьба
1	бой-баба	1	Памба	8	способа	181	судьба
24	слаба	1	дифирамба	3	герба	1	ворона-
36	раба	7	бомба	1	серба		судьба
4	штаба	4	Комба	4	ущерба	1	похвальба
4	деба	1	апломба	2	корба	1	стрельба
154	хлеба	2	Колумба	9	губа	1	мольба
57	неба	1	тумба	1	пагуба	2	гульба
7	погреба	560	оба	1	Соллогуба	1	гоньба
4	Феба	2	худоба	2	Гекуба	49	борьба
37	служба	8	жалоба	11	клуба	138	просьба
30	дружба	52	злоба	3	груба	1	письмо-
5	тяжба	2	озноба	2	сруба		просьба
13	изба	29	гроба	1	труба	9	женильба
2	биба	1	гардероба	10	шуба	31	люба
1	Скриба	5	короба	20	рыба		
....							
73	бить	7	избить	1	дрррробить	1	нарубить
2	забить	8	прибить	1	пособить	8	грубить
4	ослабить	1	зашибить	52	оскорбить	1	нагрудить
1	набить	1	пришибить	15	сбить	1	струбить
8	грабить	1	ошибить	8	отбить	1	срубить
1	заграбить	2	долбить	97	убить	3	трубить
17	отграбить	1	добить	20	губить	2	выбить
1	пограбить	1	разжалобить	5	загубить	387	любить
1	вбить	2	озлобить	39	погубить	2	залюбить
1	подбить	1	знобить	5	сгубить	1	влюбить
8	перебить	2	побить	1	усугубить	5	разлюбить
4	теребить	5	дробить	1	голубить	4	возлюбить
62	употребить	3	раздробить	1	приголубить	2	долюбить
25	истребить	2	коробить	6	рубить	67	полюбить
23	разбить	4	пробить	1	зарубить		

Не приходится разъяснять эвристическую ценность обратного словаря. Наряду с обратным словарем графических слов в электронной части Словаря представлен и обратный словарь лемм (43577 разных лемм). Весьма объемные обратные словари отражают весь корпус текстов; что касается отдельных частей корпуса, то для них будут даны сведения о частоте словообразовательных

<sup>2</sup> Все таблицы «Введения» нумеруются с начальным нулем.



х

элементов (префиксов и суффиксов). Во всех остальных таблицах принят прямой алфавитный порядок.

В существующих частотных словарях до половины общего объема приходится на ранговые словари, т. е. на таблицы, в которых единицы расположены в порядке уменьшения их частоты ( $f$ ) и соответствующего возрастания их ранга ( $r$ )<sup>3</sup>. Примером может служить таблица 3.

В описываемом Словаре ранговые словари занимают очень скромное место — в электронной части Словаря даются четыре списка по 500 самых частых графических слов для всего корпуса текстов, для совокупности художественных текстов, для публицистики и для писем; аналогичным образом будут включены таблицы слов (лемм). Такое решение объясняется просто: ранговыми словарями практически нельзя пользоваться. В них можно ответить на такие экзотические вопросы, как «какие именно слова имеют частоту 15?» или «какое слово занимает 305-е место в ранговом словаре?», но нельзя найти конкретные слова средней и низкой частоты. Если же читателю все-таки понадобится перейти от частоты к соответствующему рангу, это можно будет сделать при помощи таблицы, уместившейся на одной-двух страницах (см. таблицу 10). Структура этих кратких таблиц описана в следующем параграфе.

Все примеры таблиц, представленные до сих пор, содержат абсолютные частоты лингвистических единиц. Их преимущество — представление полного объема информации, их недостаток — сложность непосредственного сравнения данных, входящих в разные столбцы. Как правило, столбцы отражают данные разных подкорпусов, каждый из которых не совпадает по объему с другими. Например, в таблице 1 общий объем «Критики и писем» примерно совпадает, но «Художественная литература» превышает их в три с половиной раза. Конечно, рассматривая строки с небольшой совокупной частотой, читатель мысленно учтет это обстоятельство и сделает правильный вывод. Вот три примера из таблицы 1:

	Всего	Х	К	П
бритва	36	23	12	1
брошюра	25	3	13	9
брюнетка	15	14		1

Без каких бы то ни было сложных вычислений читателю ясно, что слово *бритва* сосредоточено в критике, что слово *брошюра* крайне редко появляется в художественной литературе, а слово *брюнетка* именно в этом жанре и сосредоточено. Если же совокупная частота велика, то требуется проводить некоторые арифметические операции, что без калькулятора делать трудно.

Разрешить данную трудность можно при помощи таблиц относительных частот, где частоты приведены к общему знаменателю (скажем, на 100 тыс. словоупотреблений). Именно этот принцип характеризует таблицу 2.<sup>4</sup>

Представление результатов в виде относительных частот имеет одно ограничение — оно бессмысленно в приложении к редким явлениям. В связи с этим в Словарь вводится еще и специальная мера оценки статистической значимости реальных частот:

$$S = (f - m - 1) / \sqrt{m},$$

где  $f$  — наблюдаемая частота данного события,

$m$  — математическое ожидание этого события, подсчитанное на основе какой-то нулевой гипотезы.

Эта величина нашла в Словаре самое широкое применение. Важно, что при этом в круг анализа вовлекаются также хотя и редкие, но значимые события, иногда даже двукратное появление слова или словосочетания. Предположим нам

<sup>3</sup> У многих лингвостатистиков именно ранговый словарь именуется «частотным словарем», для второго основного варианта частотного словаря они используют термин «алфавитно-частотный словарь».

<sup>4</sup> Как это принято в статистике, в таблицах относительных частот численные значения меньше 0,5 показаны многоточием.