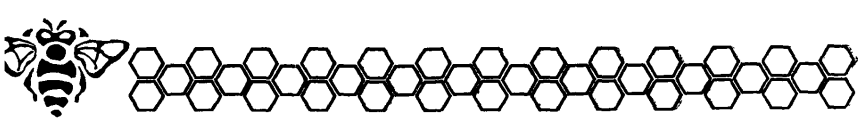

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ РУССКОГО ЯЗЫКА им. В. В. ВИНОГРАДОВА

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

СТАТИСТИЧЕСКИЙ СЛОВАРЬ ЯЗЫКА ДОСТОЕВСКОГО

ЯЗЫКИ СЛАВЯНСКОЙ КУЛЬТУРЫ
Москва 2003

ББК 83
Ш 12

Работа выполнена в рамках федеральной целевой программы «Русский язык» в 2003 году

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

А. Я. Шайкевич, В. М. Андрющенко, 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 Андрющенко
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@mtu-net.ru

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

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.

ISBN 5-94457-145-4
9 785944 571458 >

Оглавление

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 явится 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 [Spevack]. 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 корпусы 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 Кэррола [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[1] (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.

---

[1] 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*.

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 *во́ды*, *брат* и *братья* и т. п.), о формах императива многих глаголов (*не беспокойтесь*, *ступай* и т. п.). Часто различаются субстантивированные прилагательные (*артельный*, *блаженный*, *ближний*, *больной*, *большие*, *былое*, *дворовый*, *знакомый*, *рабочий* и т. п.) и омонимичные исходные прилагательные. Выделены и многие адъективированные причастия, например *благоухающий*, *верующий*, *воинствующий*, *волнующий*, *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*, *быль 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*, *Адрианополь* (город), *«Адрианополь»* (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	X	К	П
Бог	1730	1081	173	476
Бог в помочь	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 помочь*, *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, лексемы (как в таблице 001)[2], 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	женитьба
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	нагрубить
1	набить	1	пришибить	15	сбить	1	сгрубить
8	грабить	1	ошибить	8	отбить	1	срубить
1	заграбить	2	долбить	97	убить	3	трубить
17	ограбить	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

[2] Все таблицы «Введения» нумеруются с начальным нулем.

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)[3]. 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, умещающейся 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.[4]

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

---

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

[4] Как это принято в статистике, в таблицах относительных частот численные значения меньше 0,5 показаны многоточием.

надо оценить статистическую значимость слова *деньги* в макрожанре «Письма». Частота слова *деньги* в текстах Достоевского равна 2097, в письмах оно встретилось 822 раза. Предположим, что слово *деньги* не зависит от специфики макрожанра, тогда, зная долю писем во всем корпусе (0,1833), мы можем подсчитать математическое ожидание появления этого слова в письмах:

$$2097 \times 0{,}1833 = 284.$$

Подставляя величины 822 и 284 в нашу формулу, получаем S=31 (величину исключительно высокую), отсюда вывод — слово *деньги* очень характерно для писем Достоевского. Будем называть такие единицы **лексическими маркерами** подкорпусов (макрожанров, микрожанров, периодов творчества, отдельных текстов, отдельных персонажей и т. п.). Соответствующие списки целиком входят в Словарь. В этих списках частота (f) подчинена мере статистической значимости (S). Примером могут служить таблицы 14-16.

Формула оценки статистической значимости может быть использована для выявления текстуальных связей слов. Весь текст механическим образом членится на фрагменты равной длины (скажем, 40 слов), а затем подсчитывается число фрагментов, в которых одновременно встретились слово *x* и слово *y*. Если реальная частота совместной встречаемости статистически значима, делается вывод о текстуальной связи двух слов. Таким образом, в Словаре найдет отражение еще один лингвистический объект — **текстуальные связи** слов.

Так, редкое слово *агония* встретилось в жанре критики всего 4 раза, но показало текстуальные связи с пятью словами: *актер* (S=18), *естественный* (S=6), *зритель* (S=10), *правда* (сущ.) (S=2), *умирать* (S=6). Из этих пяти связей одна (со словом *умирать*) может считаться общезначимой для русского языка, остальные — обусловлены конкретным текстом, где ведется речь об изображении агонии на сцене.

Более сбалансированное соотношение общезначимых и текстуальных связей обнаружим в жанре «Критика и публицистика» у слова *Гоголь*.

Таблица 002

Текстуальные связи слова *Гоголь* (f=47)

S	f		S	f		S	f	
18	3	«Женитьба»	4	3	исчезнуть	2	2	выражаться
16	5	перевод	4	2	по-французски	2	3	господин
15	8	Тургенев	4	4	повесть	2	2	комический
13	3	«Мертвые души»	4	2	Поприщин	2	5	литература
12	5	Островский	4	4	произведение	2	4	начинать
11	2	Виардо	4	3	сверх того	2	3	писатель
11	2	непереводимый	4	5	язык	2	2	следовать
7	3	жанр	3	2	где-то	2	2	сочинение
7	3	комедия	3	2	драма	2	7	тогда
7	3	перевести	3	2	Лермонтов	2	7	хотя
7	2	Репин	3	2	портрет	2	3	художественный
5	2	Писемский	3	2	правый	2	4	что-то
5	4	смех	3	5	Пушкин	2	2	Щедрин
4	2	Диккенс	3	4	французский			

Здесь мы обнаруживаем текстуальные связи, которые могли бы ожидать от имени *Гоголь* даже не проводя специальных исследований, с другой же стороны, встречаемся со связями, обусловленными конкретным текстом, в котором речь идет о непереводимости Гоголя (*перевод, непереводимый, перевести, по-французски, французский, Виардо, Тургенев*). Подробнее об анализе текстуальных связей см. ниже.

4. Описание отдельных таблиц книги

Таблица 1 «Распределение лексем по основным жанрам» содержит 43577 лемм (и 3445 других лексем). Основные жанры (или макрожанры) сокращенно записаны как Х — Художественная литература, К — Критика и публицистика, П — письма. Для экономии места слова с частотой 1 напечатаны в подбор в конце каждой страницы, два последних жанра записаны здесь кодами К и П, слова, встретившиеся в художественной литературе, не сопровождаются каким-либо символом.

Таблица 2 «Относительная частота лексем в основных жанрах» содержит 3000 лексем. Относительные частоты округлены до целого числа, лексемы,

встретившиеся в данном жанре с частотой менее 0,000005, показаны многоточием. Следует помнить, что Средняя, указанная в первом столбце, это — средняя взвешенная, т. е. получена с учетом неравного объема трех макрожанров.

Таблица 3 «100 самых частых лемм в текстах Достоевского» представляет собой верхнюю часть рангового словаря. Символы r и f читаются как «ранг» и «частота» соответственно. Аналогично устроены таблица 4 «100 самых частых лемм в художественных произведениях Достоевского», таблица 5 «100 самых частых лемм в критике и публицистике Достоевского», таблица 6 «100 самых частых лемм в письмах Достоевского».

Большей наглядностью для читателя обладают таблицы, в которых верхушка рангового словаря представлена в распределении по знаменательным частям речи. Это таблица 7 «40 самых частых существительных», таблица 8 «40 самых частых глаголов», таблица 9 «40 самых частых прилагательных». В этих таблицах каждый столбец посвящен своему основному жанру, внутри столбца слова упорядочены по убыванию частоты.

Информация о ранговом словаре (полностью представленном в электронной части) завершается четырьмя таблицами, в которых фигурируют только числа, Это таблица 10 «Частотный спектр рангового словаря лемм всего корпуса текстов», таблица 11 «Частотный спектр рангового словаря лемм беллетристики», таблица 12 «Частотный спектр рангового словаря лемм критики и публицистики», таблица 13 «Частотный спектр рангового словаря лемм писем». Четыре столбца этих таблиц интерпретируются следующим образом: в первом столбце показан ранг леммы в ранговом словаре. Цифры в последней строке первого столбца указывают на число разных лемм — 43577 во всем корпусе, 34257 в художественной литературе, 21448 в критике и публицистике, 17367 в письмах. Во втором столбце показана частота слова данного ранга, в третьем — накопленная частота всех слов с данным рангом и ниже. Последняя строка этого столбца — совокупная частота лемм данного корпуса. Заметим, что эта цифра несколько меньше числа графических слов, поскольку среди лемм довольно много словосочетаний. Наконец, в четвертом столбце дается накопленная относительная частота, т. е. результат деления накопленной частоты данного ранга на совокупную частоту лемм.

Следующие три таблицы показывают лексические маркеры основных жанров. Это таблица 14 «Лексические маркеры беллетристики», таблица 15 «Лексические маркеры критики и публицистики», таблица 16 «Лексические маркеры писем». Лексические маркеры поданы здесь в алфавитном порядке. Таблицы 17, 18 и 19 содержат информацию о лексических маркерах, упорядоченных по убыванию S и тем отличающихся от таблиц 14-16. В кратком виде эта информация представлена в таблице 003.

В таблицах 20 и 21 представлена грамматическая информация — редкий гость в частотных словарях. В таблице 20 даются абсолютные частоты во всем корпусе и по трем основным жанрам; в таблице 21 та же информация подана в виде относительных частот.

Грамматические классы включают традиционные части речи и некоторые подклассы внутри частей речи. Особо выделены деадъективные наречия, изменившие свои синтаксические функции: примером могут служить лексемы *довольно, следовательно, действительно, точно, давно, конечно*. Русские суффиксы соотнесены с частями речи и с грамматическими подклассами внутри частей речи. Так они и сгруппированы в данных таблицах. Что касается префиксов, то они хорошо коррелируют с мотивирующими основами в процессе словопроизводства, но результирующие производные слова (часто осложненные суффиксами) уже прямо не соотнесены с частями речи. Префиксы и первые компоненты сложных слов поданы в общем алфавитном порядке.

Поскольку в префиксальных образованиях часто наблюдается процесс морфологического опрощения, особо выделяются группы слов с этимологическим префиксом, чья семантическая мотивация частично или полностью затуманена. Примером может служить группа слов с начальным в-: *вкус, влияние, вменять, вместо, вовсе, внедрить, внезапно, вникнуть, внимание, вонзить, вперить, впечатление, впиться, вплоть, впросак, вряд, всадник*. Такие группы слов даются в таблицах как отдельные строки, причем префикс в таком случае сопровождается знаком звездочки (*).

Важнейшие лексические маркеры основных жанров

Художественная литература		Критика и публицистика		Письма	
S		S		S	
44	он	74	всё	72	письмо
28	она	62	народ	70	писать
25	-с	50	мы	59	Достоевский
24	князь	50	наш	58	мой
22	вдруг	50	русский	53	роман
21	был	49	Европа	52	написать
20	я	37	народный	48	рубль
17	Алеша	36	Франция	47	Ф.
16	Лиза	35	Россия	46	получить
16	лицо	34	война	45	я
16	рука	31	европейский	42	ты
15	глаз	30	лишь	41	Аня
15	так	29	русский (сущ.)	38	Петербург
14	дверь	29	славянин	38	твой
14	комната	23	идея	37	будет
14	проговорить	23	литература	37	Паша
13	вскричать	23	новый	35	выслать
13	да (утверждение)	23	общество	34	прислать
13	Митя	23	политический	33	ваш
13	смотреть	23	статья	33	лист
13	что-то	23	турок	33	многоуважаемый
12	будто	22	восточный	32	Катков
12	весь	22	господин	31	год
12	как	22	искусство	31	деньги
12	как будто	22	он	30	если
12	какой-то	21	большинство	30	Старая Русса
12	нет (отрицание)	21	настоящий	29	целовать
12	Раскольников	21	нация	28	до свидания
12	с	21	свой	28	журнал
12	стоять	20	в	28	Николай
12	так и	20	Германия	28	просьба
12	что (местоимение)	20	идеал	28	уведомить
11	Голядкин	20	именно	27	жить
11	Грушенька	20	исторический	27	редакция
11	давеча	20	правительство	27	«Русский вестник»
11	опять	20	развитие	27	Стелловский
11	пред	20	столь	27	Федор
11	старик	20	цивилизация	27	Федя
10	арестант	20	этот	26	август
10	было (частица)	19	век	26	адрес
10	быстро	19	великий	26	«Заря»
10	глядеть	19	вера	26	о
10	говорить	19	высший	26	обнимать
10	голова	19	люди	26	Эмилия
10	заметить	19	маршал	25	брат
10	как бы	19	обвинение	25	июль
10	спросить	19	республика	25	просить
10	стол	19	факт	25	ради Бога
10	убить	19	явление		

Книга завершается тремя большими таблицами, где для основных жанров показана лексическая информация в связи с отдельными периодами жизни и творчества Достоевского. По своей структуре эти таблицы аналогичны таблице 2, отличаясь от нее содержательной интерпретацией столбцов.

В таблице 22 «Распределение лексем в беллетристике по периодам творчества» представлены пять столбцов — суммарный и четыре частных по периодам: 1844-1849, 1856-1865, 1866-1872, 1873-1880. Обычно выделяются три периода, но в данной таблице третий период представлен в виде двух столбцов, что позволит проверить устойчивость тех или иных хронологических тенденций.

Таблица 23 посвящена критике и публицистике по периодам творчества. Три выделенных периода (1845-1848, 1860-1865, 1873-1881) разделены длительными периодами публицистического молчания и в особой аргументации не нуждаются.

Заметим, что здесь исключены из рассмотрения обширные цитаты (всего 32 тыс. словоупотреблений).

Наибольшее число столбцов представлено в таблице 24, посвященной письмам Достоевского: 1832-1843 — годы учения, 1844-1849 — литературная деятельность вплоть до ареста, 1854-1859 от выхода из каторги до возвращения в Петербург, 1860-1866 литературная деятельность в Петербурге, 1867-1871 женитьба на А. Г. Сниткиной и пребывание за границей, 1872-1881 — от возвращения в Россию до конца жизни.

Объем хронологических подкорпусов отражен в таблице 004.

Таблица 004

Объем хронологических и жанровых подкорпусов
(в тысячах графических слов)

Художественная литература		Критика и публицистика		Письма	
				1832-1843	18
1844-1849	247	1845-1848	12	1844-1849	28
1856-1865	406			1854-1859	67
		1860-1865	140	1860-1866	79
1866-1872	668			1867-1871	144
1873-1880	515	1873-1881	373	1872-1881	194
Итого	1835		524		531

В следующем параграфе будут отмечены и другие таблицы, представленные только в электронной части Словаря.

5. Перспективы использования Словаря

Уже из предыдущего изложения должно быть ясно, что основной путь получения все более детальной информации заключается в последовательной дифференциации совокупного корпуса текстов Достоевского. В таблицах книжной части Словаря более или менее полно нашли отражение два верхних уровня получающейся иерархии подкорпусов: основные жанры (или макрожанры) и хронологические подкорпусы внутри макрожанров. На каждом таком уровне можно исследовать статистически значимые расхождения между подкорпусами в попытке выявить характерные черты каждого из подкорпусов и определить их взаимоотношения. Попробуем самым кратким образом показать возможности этого пути.

Обратимся для начала к характеристике основных жанров при помощи таблиц 1-21. Некоторые из этих таблиц (например, таблицы 10-13) могут показаться совершенно эзотерическими, сама интерпретация их потребовала бы слишком много места для нашего «Введения». Но даже такие таблицы могут дать полезные результаты. Например, сравнение таблиц 12 и 13 ясно показывает, что при равном объеме корпус критики и публицистики и корпус писем сильно расходятся по числу лемм: 21448 разных лемм в критике и публицистике,

17367 — в письмах. Еще больше расхождения в числе разных графических слов (соответственно 59446 и 43689)[5].

Обращение к ранговым словарям (таблицы 4–9) кажется и убедительным, и наглядным. Действительно, слово *человек* возглавляет список существительных в художественных произведениях, оно занимает 4-е место в критике и публицистике и лишь 19-е в письмах. Однако следует помнить, что небольшие изменения в частоте могут сильно сказываться на месте в ранговом списке. Таблица 2 показывает следующие цифры относительной частоты (на 100 000 слов) — 219 в художественной литературе, 196 в критике и 108 в письмах. Противопоставленность писем двум остальным жанрам (в отношении слова *человек*) можно считать доказанным.

Для взвешенного суждения о противопоставлении подкорпусов должны использоваться таблицы всех трех типов: таблицы относительных частот, ранговые словари и таблицы лексических маркеров. При этом надо учитывать, что отдельные слова и целые классы слов обладают разной дифференцирующей силой. Так, в таблице 7 мы находим всего 10 существительных, присутствующих во всех трех списках (*человек, дело, время, раз, день, сердце, жизнь, случай, мысль, год*). Пять слов (*минута, люди, лета, вопрос, душа*) объединяют списки художественной литературы и критики; семь слов представлены в первом и третьем списках (*письмо, Бог, ночь, брат, вид, друг, час*). Всего два слова (*журнал и статья*) объединяют второй и третий списки. Таблица 8 содержит 17 глаголов, общих всем трем спискам (*быть, знать, говорить, мочь, сказать, хотеть, стать, видеть, думать, любить, пойти, взять, прийти, сделать, иметь, жить, дать*). Отсюда следует вывод — в сравнении с существительными глаголы в меньшей степени различают жанры.

Таблица 21 особенно интересна при поиске межжанровых различий.

Художественные произведения характеризуются повышением доли глаголов (особенно глаголов на *-ся*), деадъективных наречий, местоимений-существительных, местоимений-наречий, местоимений 3-го лица, междометий, слов *да* и *нет*, частиц при местоимениях (типа *какой-то, какой-нибудь*). Среди аффиксов отметим уменьшительные суффиксы, особенно слов женского рода (*-к-а, -очк-а, -шк-а*), суффикс прилагательных *-ив-ый*, глагольные суффиксы *-ну-ть* и *-ива-ть*, продуктивные префиксы *воз-, за-, по-, под-, раз-* и *у-*. Ярких отрицательных маркеров у этого макрожанра нет.

Жанр критики и публицистики очень выразительно характеризуется грамматическими показателями. Яркими положительными маркерами выступают имена нарицательные (особенно среднего рода), прилагательные, суперлативы, местоимения-прилагательные, союзы, вводные слова. Два последних класса указывают на синтаксическую сложность. В отличие от художественной литературы с ее преобладанием глаголов, данный жанр носит именной характер. Глаголы оказываются отрицательными маркерами. Тяга к обобщениям и универсальности косвенно проявляется в деперсонализации, местоимения 1-го и 2-го лица становятся отрицательными маркерами жанра. Среди аффиксов характерны суффиксы абстрактных существительных: *-ость, -ство, -ние, -тие*; префиксы *без-* и *не-*; сложные слова с первыми компонентами *едино-, обще-, противо-, само-*. Очень характерны заимствованные суффиксы: *-аж, -изм, -ент, -мент, -ика, -ема, -ура, -ия*.

Положительных маркеров жанра писем очень немного, это — числительные, имена собственные, местоимения 1-го и 2-го лица, модальные слова. Яркие отрицательные маркеры — возвратные местоимения и местоимения 3-го лица. Обращение к таблицам лексем[6] дополнит эту общую характеристику жанра писем еще

---

[5] Адепты лингвостатистики обычно интерпретируют подобные различия в терминах богатства и бедности словаря. Лингвисты должны быть осторожнее: им нужно сначала договориться о значении терминов «богатство» и «бедность». Действительно ли наличие трех разных лемм *анализ, анализировать, анализирование* в жанре критики и публицистики свидетельствует о большем богатстве в сравнении с жанром писем, где представлено только слово *анализ*? Примем ли мы как контраргумент три разные леммы *безденежно, безденежный, безденежье,* встретившиеся в корпусе писем?

[6] Лексические маркеры берутся из таблиц 17 (художественная литература), 18 (критика и публицистика) и 19 (письма). В квадратных скобках приводятся лексические маркеры с меньшими значениями S, взятые соответственно из таблиц 14, 15 и 16.