STUDIA PHILOLOGICA

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

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

ДИСТРИБУТИВНО-СТАТИСТИЧЕСКИЙ АНАЛИЗ ЯЗЫКА РУССКОЙ ПРОЗЫ 1850—1870-х гг.

Том 3

Издательский Дом ЯСК
Москва 2021

УДК 811.161.1
ББК 81.2 Рус
Ш 17

Рецензенты:
д. ф. н. А. Ф. Журавлев,
д. ф. н. Л. Л. Шестакова

Утверждено к печати ученым советом
Института русского языка им. В. В. Виноградова РАН

Ш 17 **Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.**

Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. Т. 3. — М.: Издательский Дом ЯСК, 2021. — 768 с. — (Studia philologica.)

ISBN 978-5-907290-61-7

Том 3 монографии посвящен анализу среднего интервала текста. Исследуемый корпус прозы автоматически членится на фрагменты по 40 слов. Если совместная встречаемость двух слов во фрагментах существенно превышает величину, подсчитанную на основе нулевой гипотезы, делается вывод о наличии связи между этими словами. В результате возникает огромная сеть текстуальных связей слов. Показан способ выявления кластеров в этой сети.

Полностью сеть текстуальных связей слов (26 тысяч слов и 500 тысяч связей) представлена на компакт диске.

УДК 811.161.1
ББК 81.2 Рус

ISBN 978-5-907290-61-7

Часть 6

Текстуальные связи слов в общем корпусе русской прозы 1850—1870-х гг.

Напомним основные положения нашей монографии.

Дистрибутивно-статистическим анализом (ДСА) называем формальный (без обращения к смыслу) метод изучения больших собраний текстов, опирающийся исключительно на статистику распределения графических слов. В данной монографии этот метод прилагается к корпусу русской прозы 1850–1870-х гг. общим объемом более 14 миллионов словоупотреблений.

Вводимое понятие **интервала текста** позволяет надеяться на получение самых разнообразных содержательных результатов. В опубликованных томах монографии были исследованы интервалы, задаваемые исходным членением текста. В Части 2 (т. 1) исследовалась комбинаторика букв внутри графических слов (**микроинтервал**), в Части 4 (т. 2) объектом анализа были бинарные словосочетания (**минимальный интервал**). Эти интервалы приводили к открытию грамматики.

Остальные интервалы задаются исследователем. В рамках заданного интервала корпус текстов автоматически членится на фрагменты равной длины, каждому фрагменту присваивается свой адрес. Для пары слов определяется число общих адресов; если оно существенно превосходит математическое ожидание, подсчитанное в предположении независимости слов, делается вывод о том, что между этими словами обнаружена **текстуальная связь**. Мерой неслучайности связи служит выражение

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

где x — наблюдаемое число общих фрагментов, а m — математическое ожидание. Все предварительные исследования показывают, что, начиная с $S = 3$, результаты оказываются осмысленными. Третий (заключительный) том монографии базируется на **среднем интервале**.

6.1. Средний интервал

Перед окончательным выбором длины фрагмента для среднего интервала было проведено пробное исследование на материале текстов Тургенева (720 тысяч словоупотреблений). Сравнивались результаты при длине фрагмента в 40 слов, в 200 слов и в 1000 слов. В первом случае слово *дуэль*, например, обнаружило связь с *драться* ($S = 15$), при длине в 200 слов с *дуэлью* связаны ($S > 4$): *вызов*, *драться*, *пистолет*, *поединок*, *противник*, *пуля*, *секундант*. Значение меры неслучайности (S) напрямую зависит от объема изучаемого корпуса. Все семь слов, ассоциированных с *дуэлью* у Тургенева, будут обнаружены на всем корпусе прозы уже при длине 40 слов, к ним присоединятся еще 47 слов — *бретер*, *вызвать*, *выстрелить*, *извиниться*, *Лермонтов*, *обида*, *оскорбить*, *отказаться*, *пощечина*, *разжаловать*, *ранить*, *скандал*, *соперник*, *стреляться*, *трус*, *убить*, *удовлетворение*, *целить* и т. п. Переход к длине фрагмента в 1000 слов покажет уменьшение значений S у слов, обозначающих стандартизованные ситуации вроде дуэли; напротив, появятся сюжетные текстуальные связи, приуроченные к конкретному тексту (*болгарин — Кунцево*, *резать — лягушка — нигилист*). Это свидетельствует о качественно ином — **большом интервале**. Что касается среднего интервала, то в качестве окончательной была принята длина фрагмента в 40 слов.

Для конкретных пар слов результаты выдаются компьютером в следующем виде:

```
волосы густой    195   3156   1687   48   13.90
лежать постель   281   5303   2142   45   29.66
лес куст          99   3692   1022   28    9.85
```

Левый столбец показывает число общих адресов двух слов, во втором столбце дается частота (число адресов) первого слова, в третьем — частота второго слова, в четвертом столбце указывается величина S[1], в пятом — величина m.

Именно в таком виде отражены итоги в грандиозной табл. 6-С, представленной в электронной версии тома. Эта таблица включает более 26 тысяч слов и примерно 500 тысяч текстуальных связей. В табл. 6-С вошли слова с относительной частотой 2 и более на 1 миллион словоупотреблений, при этом учтены пары слов, имеющие по крайней мере 3 общих адреса.

[1] Величина S округляется до меньшего целого числа.

Даже двукратное совместное появление может свидетельствовать о вполне реальных семантических связях, ср.

жмых масло	2	3	642	14	0.01
зиждиться основание	2	12	932	5	0.03
калашный калач	2	13	159	13	0.01
загородный проспект пять углов	2	6	10	79	0.00[1]
индийский океан доброй надежды	2	4	33	53	0.00
казачок отплясывать	2	11	18	43	0.00.

Слово *жмурки* по два раза встретилось со словами *вздумать*, *давай*, *колокольчик*, *побежать*, *прятаться*, *разобрать*, *ребята*, *сзади*. Его текстуальные связи с числом общих адресов 3 и более будут отражены в табл. 6-С — *бегать*, *глаз*, *завязанный*, *игра*, *играть*, *ловить*, *товарищ*, *угол*.

Как видим, средний интервал во многом наследует результаты минимального интервала, ср. пару *волосы — густой*, пары *ахиллесов пята* ($S = 170$), *выеденный яйцо* ($S = 129$), *несолоно хлебать* ($S = 123$), *отложной воротничок* ($S = 132$). Пару *лежать — постель* можно было бы обнаружить на **малом интервале** (длина фрагмента 3—5 слов), ср. также пары *пазуха — вынуть*, *скрестить — грудь*, *ухаживать — больной*, *шесток — сверчок*, *юг — север*[2]. Однако большинство найденных пар обнаруживается именно на среднем интервале: не только *лес — куст*, но и *лес — дерево*.

У слова *волосы* обнаружено 912 текстуальных связей, из них 230 имеют $S = 3$, т. е. на самом пороге статистической значимости, но и в этой группе находим слова, явно связанные с волосами (*брюнет*, *дергать*, *копна*, *кружок*, *лезть*, *лен*, *немытый*, *непокорный*, *отвиснуть*, *погладить*, *причесываться*, *пушок*, *разглаживать*, *редеть*, *убранный*). Унаследованные от меньших интервалов связи нашего слова доминируют в группе максимальных значений S: начиная со связи со словом *прядь* ($S = 87$) и далее по мере уменьшения S: *седой*, *дыбом*, *ерошить*, *русый*, *черный*, *вьющийся*, *густой*, *проседь*, *зачесанный*, *каштановый*, *белокурый*, *длинный*, *расчесывать*, *растрепанный*, *всклокоченный*, *распущенный*, *причесанный*, *смоль* ($S = 31$). Уже при $S = 47$ появляется первое пополнение среднего интервала — слово *лицо*, за которым следуют *борода*, *лоб*, *висок*, *глаза*, *рост*, *голова*, *губы*, *затылок*, *карий*, *голубой*, *нос* ($S = 20$). Постепенно проясняется весьма характерная для прозы совокупность слов, описывающих внешность человека.

6.2. Лемматизация для среднего интервала

В частях 2 и 4 монографии объектами подсчетов были графические слова, в ходе работы ДСА возникали некоторые парадигмы, в общем совпадающие с традиционными парадигмами русской грамматики. Как важнейшее расхождение отметим то обстоятельство, что формы на *-о* (типа *хорошо*) целиком входили в парадигму адъективов (*хороший*) независимо от их синтаксической функции.

Опираясь на результаты частей 2 и 4, для каждой парадигмы мы объявляли леммой одно из графических слов парадигмы. Такой выбор был максимально приближен к грамматической традиции: для глаголов ею служила форма инфинитива, для субстантивов и числительных форма им. п. ед. ч. и т. п. Далее в третьем томе, говоря о словах, мы будем иметь в виду леммы.

Лемматизация для среднего интервала заставила нас в какой-то степени отойти от строгого формализма ДСА. Некоторые нерегулярные парадигмы были достроены до конца, не дожидаясь формального эмпирического подтверждения. Леммы *лететь* и *хотеть* выбирались для двух частично различающихся парадигм (*летит* и *хочет*). Парадигмы глаголов *стать* и *деть* включают формы с *-н-* (*станет* и *денет*); считаются одной парадигмой формы супплетивного глагола *идти* и его производных. Совершенно уникальная парадигма глагола *есть* (*ем*, *ест*, *едят*, *ел* и т. д.) принимается без доказательств, а его лемма отделяется от *есть* глагола *быть*.

Неформальным способом (т. е. с опорой на знание языка человеком) были разведены многие лексикографические омонимы, потенциально распознаваемые с опорой на дистрибутивные свойства. Такие омографические слова часто представлены целыми сериями, например, совпадающими формами им. п. ед. ч. субстантивов женского рода и род. п. ед. ч. субстантивов мужского и среднего рода (*раба*, *супруга*, *чада*, *ложа*, *маркиза*, *кружка*, *математика*,

[1] Компьютерная выдача 0,00 указывает на $m < 0{,}01$.

[2] Упоминание малого интервала основывается здесь скорее на общих соображениях, нежели на эмпирических данных. Проверка ДСА на малом интервале потребовала бы слишком больших затрат человеческих сил и машинных ресурсов, что сильно затормозило бы переход к среднему интервалу, обещающему больше нетривиальных результатов.

критика, *политика*, *практика*, *стрелка*, *полка*, *дымка*, *банка*, *дока*, *тока*, *треска*, *соска*, *глотка*, *лука*, *сука*, *казачка*, *скачка*, *сучка*, *душка*, *пушка*, *пенька*, *пола*, *корма*, *кума*, *вина*, *ворона*, *жара*, *пара*, *манера*, *леса*, *черта*, *уха*). Омографически совпадают формы им. п. ед. ч. субстантивов женского рода и инфинитивов глаголов: *подать*, *знать*, *стать*, *пасть*, *напасть*, *почесть*, *течь*, *печь*. Неформальным образом разведены семантически связанные адъективы и субстантивы: *легкое*, *жаркое*, *животное*, *молодые*, *данные*, *присяжные*, *понятые*, *рабочий*, *мировой*, *часовой*, *золотой*, *малый*, *блаженный*.

Учет синтаксических и дискурсивных функций побуждал нас изымать одну из форм парадигмы в отдельную лемму. У субстантивов эта операция проделана над формами творительного падежа: *порой*, *словом*, *бегом*, *мигом*, *кругом*, *задом*, *градом*, *следом*, *разом*, *образом*, *толком*, *боком*, *залпом*, *даром*, *верхом*. У адъективов отдельной леммой стали предикативы: *прав*, *каков*, *таков*, *здоров*, *готов*, *мал*, *убежден*, *принужден*, *вынужден*, *сужден*, *виден*, *должен*, *сложен*, *нужен*, *болен*, *волен*, *наклонен*, *покоен*, *уверен*, *намерен*, *согласен*, *известен*, *грешен*, а также *главное*. У глаголов отдельными словами стали формы императива: *посуди*, *извини*, *давай*, *ступай*, *послушай*, *прощай*, *постой*, *здравствуй*, *помилуй*, *оставь*, *изволь*, *берегись* и (только формы ед. ч.) *кажись*, *поди*, *пускай*. К ним присоединяются и некоторые другие формы: *положим*, *бывало*, *стоило*, *следует*, *надлежит*, *хватит*, *значит*, *было* (частица), *кажется*, *казалось*, *признаюсь*, *будет* (= хватит), *разумеется*.

Главным мотивом выделения адъективных форм на *-о* был статистический фактор: *много*, *далеко*, *жалко*, *хорошенько*, *мало*, *давно*, *слезно*, *чрезвычайно*, *особенно*, *непременно*, *совершенно*, *нечаянно*, *охотно*, *известно*, *совестно*, *нарочно*, *скоро*; иногда учитывалась и дискурсивная функция: *право*, *здорово*, *бывало*, *ровно*, *ладно*, *видно*, *угодно*, *стыдно*, *должно*, *возможно*, *нужно*, *полно*, *собственно*, *верно*, *согласно*, *точно*, *действительно*, *относительно*, *больно*, *довольно* — всего 47 случаев.

Преимущественно семантические соображения доминировали при разделении на леммы омонимичных, омографичных или полисемичных субстантивов. Примерами могут служить: *пол* (floor/gender), *замок* (lock/castle), *месяц* (month/moon), *лавка* (shop/bench), *лицо* (face/person), *свет* (light/world/society), *образ* (image/icon), *ключ* (key/spring), *глава* (chapter/head), *присутствие* (presence/office), *икра* (caviare/calf), *среда* (milieu/Wednesday), *мука* (torment/flour), *мир* (world/peace/community), *состояние* (state/fortune), *град* (hail/city), *пост* (fast/post), *член* (member/limb) — всего 79 случаев.

Содержательные и статистические соображения учитывались при объединении в особую лемму сочетания слов: *друг друга*, *мой друг*, *во веки веков*, *не след*, *в качестве*, *в течение*, *чуть не*, *по крайней мере*, *не в духе*, *как раз*, *молодые люди*, *в сопровождении*, *в состоянии*, *не правда ли*, *с какой стати*, *по милости*, *в особенности*, *в сущности*, *в отдельности*, *без памяти*, *по крайней мере*, *так называемый*, *молодой человек*, *правитель дел*, *коллежский советник*, *статский советник*, *тайный советник*, *милости просим*, *одним словом*, *право слово*, *чего доброго*, *не до*, *как только*, *как нарочно*, *ни за что*, *быть может*, *Невский проспект*, *слава богу*, *в пику*, *на лету*, *к лицу*, *лицом к лицу*, *не в силах*, *на днях*, *как бы*, *так сказать*, *то есть*, *стало быть*, *про себя*.

Подобные леммы в табл. 6-С исчисляются сотнями, но за редкими исключениями в табл. 6-А они не включались как основные (см. § 6.3).

Отметим два важных ограничения размера словника, подвергнутого анализу в третьем томе.

Наша основная формула, восходящая к распределению Пуассона, имеет смысл в применении к событиям с малой вероятностью. В минимальном интервале самое частое слово (*и*) имело вероятность 0,045, что оправдывало применение этой формулы ко всем графическим словам без исключения. При выбранном размере фрагмента (40 слов) у пятидесяти самых частых слов эта вероятность превосходит 0,1. Лишь два глагола из этого числа (*говорить* и *знать*) учитывались при выявлении текстуальных связей; глаголы *быть*, *сказать* и *мочь* из анализа исключались. К тому же порогу приближается вероятность двух субстантивов (*дело* и *рука*), которые мы включили в подсчеты. Вероятность самого частого адъектива (*хороший*) равна 0,045. Таким образом, в определении текстуальных связей учитывались субстантивы, адъективы и глаголы (за исключением трех, указанных выше). Включены в анализ и числительные (за исключением слов *один*, *первый* и *два*). Что касается других частей речи, то они включались в анализ спорадически и, конечно, при меньшей вероятности: *вдруг*, *всегда*, *куда*, *назад*, *чуть*, *дома*, *домой*, *чай*, *из-за*, *на* (частица), *вчера*, *вон* (прочь), *прочь*, *гораздо*, *правда* (союз), *слегка*, *вслед*, *везде*, *сквозь*, *жаль*, *замуж*, *вокруг*, *вскоре*, *вроде*, *впереди*, *спасибо*, *изредка*, *накануне*, *бишь*, *вишь* и др. Междометия учтены полностью — *ах*, *о* (!), *а* (!), *ох*, *эх*, *ай*, *ой*, *фу*, *ого*, *гм*, *ха-ха*, *хе-хе*, *хи-хи*, *увы*.

Второе ограничение касается имен собственных. В отличие от малых интервалов в третьем томе учитываются только **реальные и общеизвестные имена**: *Луга*, *Глинка*, *Москва*, *Москва-река*, *Мойка*, *Таганка*, *Фонтанка*, *Ильинка*, *Владимирка*, *Китеж*, *Северная пчела*, *Выборгская сторона*, *Разъезжая улица*, *Ревизор*, *Дездемона*, *Яго*, *Венера*, *Добролюбов*, *Грибоедов*, *Хлестаков*. Личные имена и имена с отчествами учитываются, когда они принадлежат монархам — *Алексей Михайлович*, *Екатерина*, *Павел*, *Наполеон*. Не учитываются ни имена персонажей в рамках определенного текста, ни соответствующая топонимика, ср. *Карамазов*, *Алеша*, *Ракитин*, *Дарданелов*, *Зосима*, *Смердяков*,

Трифон Борисыч, *Грушенька = Аграфена Александровна*, *Мокрое*, *Воловья*, *Скотопригоньевск* в известном романе Достоевского. Можно полагать, что наше исследование остается на 90 % формальным даже с учетом всех «семантических» отступлений, описанных в этом параграфе.

6.3. Представление текстуальных связей

Та форма компьютерной выдачи, о которой шла речь выше и которая сохранена в табл. 6-С, возможна только в электронной версии. Ее важное преимущество — возможность прямого использования при расширении корпуса текстов. Для удобства человеческого восприятия желательны и другие формы представления текстуальных связей.

Совершенно исключается визуальное представление связей в виде графа. Даже современная вычислительная техника не может отразить граф с десятками тысяч вершин и сотнями тысяч ребер, причем ребер с определенным весом (*S*).

Представление в виде матрицы возможно лишь при небольшом числе строк и столбцов. Как пример рассмотрим квадратную матрицу текстуальных связей пяти слов с их значениями *S* (слева показаны частоты соответствующих слов).

Таблица 6.1
Текстуальные связи обозначений времен года

F		год	весна	лето	осень	зима
9730	год	-	14	12	15	12
1074	весна	14	-	41	15	42
1716	лето	12	41	-	47	76
788	осень	15	15	47	-	51
1515	зима	12	42	76	51	-

Высокие значения *S* с несомненностью доказывают реальность общеязыкового кластера четырех слов и их общую связь со словом *год*[1].

Конечно, четыре слова табл. 6.1 имеют и другие текстуальные связи. Вот важнейшие из них (*S* > 10):

весна — вода, вскрыться, грач, зазеленеть, кулик, наступить, половодье, появляться, прилет, пролетный, ранний, снег, стая, теплый, утка;
лето — гриб, дача, деревня, жаркий, курорт, нынешний, провести, прошлый, черника;
осень — дождливый, дождь, наступить, отлет, охотник, поздний, пролетный, стая, стрельба, теплый;
зима — бесснежный, замерзать, зимовать, иней, корм, мерзнуть, мороз, прорубь, прошлый, снег, студеный, теплый.

Столь же реально существование общеязыкового кластера следующих четырнадцати слов.

Таблица 6.2
Текстуальные связи названий месяцев

F		г	м	я	ф	м	а	м	ин	ил	а	с	о	н	д
9730	год	-	20	18	15	23	12	19	16	15	17	15	10	15	22
3601	месяц	20	-	6	8	12	15	23	17	15	11	10	9	12	7
271	январь	18	6	-	29	20	13	7			4	6	6	14	24
262	февраль	15	8	29	-	36	13	11				3	8	10	12
311	март	23	12	20	36	-	42	18		3	4	4	14	11	3
359	апрель	12	15	13	13	42	-	28	6	9	9	5	5	8	6
496	май	19	23	7	11	18	28	-	25	14	8	4	7	5	
412	июнь	16	17	20			6	25	-	40	10	9			
348	июль	15	15			3	9	14	40	-	26	13	5		
476	август	17	11	4		4	9	8	10	26	-	40	4	12	7

[1] Заметим, что слово *сезон* (часто принимаемый в качестве гиперонима четырех слов) встречается на порядок реже (*F* = 96), но связано с соответствующими прилагательными — *зимний* (*S* = 28), *весенний* (*S* = 9), *летний* (*S* = 41), *осенний* (*S* = 12).

435	сентябрь	15	10	6	3	4	5	4	8	13	40	-	40	25	7
273	октябрь	10	9	6	8	14	5	7		5	4	40	-	36	11
297	ноябрь	15	12	14	10	11	8	5				25	36	-	23
332	декабрь	22	7	24	12	3	6					7	11	23	-

Все 12 названий месяцев связаны с гиперонимом *месяц* и со словом, обозначающим целое (*год*), максимальные значения S (> 20) расположились вдоль главной диагонали, что соответствует реальной хронологической последовательности месяцев.

Таблица 6.3
Текстуальные связи названий дней недели

	д	н	п	в	сер	ср	ч	п	с	в
день	-	27	10	12		7	7	11	11	16
неделя	27	-	11	14	4	3	13	15	9	12
понедельник	10	11	-	66	29	14	24	25	41	27
вторник	12	14	14	-		78	42	26	39	13
середа		4	29		-		32	74	8	8
среда	7	3	14	78		-	52	52	13	12
четверг	7	13	24	42	32	52	-	47	20	15
пятница	11	15	25	26	74	52	47	-	27	12
суббота	11	9	41	39	8	13	20	27	-	50
воскресенье	16	12	27	13	8	12	15	12	50	-

У слова *воскресенье* отмечена также связь со словом *будни* (S = 21). Слова *понедельник* и *вторник* связаны со словом *постный*; слова *вторник*, *середа*, *среда*, *пятница* — со словом *пост*.

В табл. 6.2 и 6.3 мы столкнулись с технической проблемой размещения столбцов на странице книги; задача была решена путем усечения названий столбцов. Такая же проблема в табл. 6.4 решена путем замены числительных на их цифровые эквиваленты.

Таблица 6.4
Текстуальные связи числительных

	3	4	5	6	10	20	30	40	50	100	200	300	400	500	1000
три	-	43	16	13	10	14	9	9	7	8			5	3	49
четыре	43	-	21	12	5	24	8	9	5	7	4	3	4		20
пять	16	21	-	42	21	107	54	39	23	18	10	8	7	9	53
шесть	13	12	42	-	10	18	11	10	7	5					19
десять	10	5	21	10	-	25	9	7	10	10	10	3		4	49
двадцать	14	24	107	18	25	-	32	16	19	27	9	4	9	6	44
тридцать	9	8	54	11	9	32	-	30	14	12	11	9		3	36
сорок	9	9	39	10	7	16	30	-	19	9	4	6	8		22
пятьдесят	7	5	23	7	10	19	14	19	-	41	50	26	12	16	52
сто	8	7	18	5	13	27	12	9	41	-	35	16	15	12	73
двести		4	10		10	9	11	4	50	35	-	46	22	24	61
триста		3	8		3	4	9	6	26		46	-	32	28	47
четыреста	5		7			9		8	12	15	22	32	-	17	23
пятьсот	3		9		4	6	3		16	12	24	28	17	-	52
тысяча	49	20	53	19	49	44	36	22	52	73	61	47	23	52	-
аршин	14	10	4	5		4					3				
ассигнация	7	5	9	5	8	5	5		21	17	12	14		15	31
билет	4		10		3		5	3	4	6		6		3	18
бумажник			4							3	3	9	5	11	6
в течение	9		10	5	3	8									
вексель			6		3	7	9	3	7	11	4		3		34
верста	22	18	39	16	33	33	32	37	25	33	18	14	13	15	19
вершок	11	14		5	6										
взаймы	5		6	4	6				6	6	4	11		5	14
восемь	7	17	14	12	17	27	20	18	6	6	5	3	4	4	21
восемьсот						7	13	7	10	5	17	4	8	10	43
выиграть					5					8	16	6	4	5	16
год	43	34	12	5	9	37	15	14	12	11	11	10		4	17
годовой	4	3				4		4		7	3			4	24
градус					3	10	11	9		3					

	3	4	5	6	10	20	30	40	50	100	200	300	400	500	1000
гривна	7	7	4	45	5			7							
гульден		7	6			7			4	7	9	6			18
дать	4		8			6	3		9	10	8	6	3	10	19
двенадцать	3	8	5	12	15	6		4						4	16
девяносто			7	5			4	3	3	15	4				12
девять	8	11	6	4	11	10	5	7	8	7		3			9
девятьсот					3		5						16		15
день	44	16	14	9	12										
деньги	11		10	4	9	11	7	4	17	20	23	23	9	15	48
десятина	5	10	6	3	7	11	9	6	4	14	15	7		7	25
десяток	10	7	3												23
длина	12	6	4	7		3									
доход	3		7			6	3		6	13	7	3		8	31
душа(*адм.*)	5	4	8	4	6	11	18	6	18	21	26	40	13	26	47
дюжина	9	6				3									
жалованье	4		8		5	4		4	9	5	7	8			16
заплатить	5	3	9	5	7	7	5	3	6	7	11	8		10	17
имение			6			6	5		3	5	3	7	4		16
итого	9	3	5	4	3	3			14		12				14
капитал			6		4	7	3		9	9	7		4		22
карман	3		6		3	4			3	5	5			4	9
копейка	15	7	29	15	17	30	27	23	26	7	7	3	5	4	15
купить	5		10		5	6			3	6	4	3		4	
лет	9	7	87	32	62	106	98	84	53	29	9	4			10
лишком	6	7		4		14	16	17	6	5	12	13	6	3	9
месяц	46	23	17	31	10	7	5		9	9		6			
миллион	4		12		6	5			10	18	14	9		11	33
миля	7		8	4	5	7		7	4	5	8	15			
минута			33		25										
минуть			9	3	4	9	5	5							
назад	18	13	19	9	14	17	10	6	3		3				
наличный				5					3		5	9			20
неделя	33	8	9	24		4									
отсчитать	3		6		4	3	3	6	9	6	5	11		10	13
платить	5	3	8	6	5	4	6		6	6		6		6	11
половина	10	9	6	6	4	3	4		3		3				9
полтина	6	8	10	6	5	6	3	5	5	3				3	
полтора	10	4		3		4						4		10	45
полтораста		3							8	26	8	5			25
получать	5		9	3	5	7	5		6	7	4	10		10	19
получить	7	4	9	4	3		3		7	6	6	6	5	5	20
пополудни	9	18	11	8	4										
продать	3		8	3	4	4		3	7	5	7	6			16
прожить	11	6	9	4	3	7	4		4	12	4				7
проиграть	6			3	8				6	8	4	3		7	22
пройти	11	5	9		10	3									
процент	4		16	3	20	14	4		9	14	5	5	4		23
пуд	12	6	12	12	7	9			5	10		6			18
пятнадцать	4		11	3	18	22	7	8	5	5		6			33
пятый		13	4			16	6	4		4					
расстояние	5				5	3	3		7	3	4				
расход			4		3	3	4	6	9	4		4			8
ровно	5	4	6	5	3										
рост			3			7	13	11	5						
рублик	8		12		4	12			6	8	7				
рубль	33	18	76	25	56	57	33	22	70	94	79	90	40	87	100
сажень	12	5	10	9	14	14	7		13	21	12	3			
семь	10	8	13	34	9	25	14	10	7	4				5	17
семьдесят	6	5	43	4		8	17		18	22	7	3			27
семьсот		3	4			5			19	3		11		13	25
серебро	12	6	20	6	10	16	10	6	17	16	13	18	12	11	47
сосчитать	4		4				4	3	5	3	7	7			13
сотня	6	3	8	3		3				5				3	42
спустя	8		9		8										

	3	4	5	6	10	20	30	40	50	100	200	300	400	500	1000
средний	7		4			3	7	5							
сряду	16	10	12	8	10	5	4	7							
стоить			9	4	7		5	4	7	6		11		7	15
сумма	6		10	3	7	6	6	3	4	5	9	5	9	5	33
сутки	6	9	5			8						5			
талер	5				8	4				11					10
тысчонка	5		9				4			6					7
уплата			6						5	6	3	3		6	19
утро	4	6	7	15	11										
флорин		18				4			6	8					35
франк	3	7	6		13	15	3		22	22	9	10	20	21	61
фунт	14	9	15	8	12	10	10	6	5	3	8				3
целковый	31	9	21	20	19	13	6	11	23	25	20	26	8	13	26
цена	26	3	8	5	4	7		6	3	6	4	4		3	15
цифра			5	4	6	6	5	3	5	6				8	18
час	29	36	27	41	36	6									
человек		3	9	6	5	11	10	8	8	4	3		3		
червонец			6							5	5	10		4	15
четвертый	12	8				12	7	3					3		
четверть	16	7	3	4		3									
число	6	4	5		6	8	4	4	5		5				5
чистоган						3	5		9						23
шаг	6		6		16	8	5		6	10	7	3		3	
шестьсот				7					7		17	9		19	25
штука	4	3	7	3	3	4	5		4	6					3
этаж	7	6	5												

Верхняя часть табл. 6.4 представляет собой квадратную матрицу пятнадцати числительных; последующие 105 строк показывают текстуальные связи числительных с другими словами. При этом расширяется самый круг числительных (*семь*, *восемь*, *девять*, *двенадцать*, *полтора*, *пятый* и др.), к нему присоединяются и другие слова с семантикой числа (*число*, *цифра*, *десяток*, *дюжина*, *лишком*, *половина*, *ровно*, *сотня*, *средний*, *сумма*). Главная сфера, ассоциированная с числительными, — сфера денег и имущества (39 строк — *деньги*, *рубль*, *копейка*, *целковый*, *франк*, *серебро*, *ассигнация*, *стоить*, *цена*, *купить*, *продать*, *взаймы*, *имение*, *десятина*, *душа* (адм.), *вексель*, *доход*, *жалованье*, *процент*, *миллион*, *тысчонка* и т. п.). Две другие важнейшие сферы, где появляются свои объекты счета, — ВРЕМЯ (*год*, *лет*, *месяц*, *неделя*, *сутки*, *день*, *утро*, *пополудни*, *час*, *четверть*, *минута*, *пройти*, *минуть*, *назад*, *спустя*) и ПРОСТРАНСТВО (*верста*, *миля*, *сажень*, *аршин*, *вершок*, *длина*, *расстояние*)[1].

В тех случаях, когда требуется показать внутреннюю неоднородность круга текстуальных связей слова, таблица приобретает более сложный вид. Примером послужит табл. 6.5, где пронумерованы ассоциаты слова *лук* и показаны их связи друг с другом. Поясним структуру таблицы на примере слова *запах*: значимость его связи с основным словом *лук* указана в скобках — в данном случае ($S = 8$); далее следуют номера слов, с которыми связано *запах* и после двоеточия дается соответствующее *S*. Мы видим, что среди ассоциатов слова *лук* слово *запах* в максимальной степени связано со словом *пахнуть* ($S = 46$).

В левой части таблицы расположились слова, так или иначе связанные друг с другом. С точки зрения «семантической» лингвистики все они ассоциируются со словом *лук* ‘овощ’. В правой части таблицы находим ассоциаты омонима *лук* ‘оружие’. Все они связаны друг с другом и никак не связаны со словами левой части.

Таблица 6.5
Текстуальные связи слова *ЛУК* ($S > 3$)

1. вонять (*S* = 11) 17:11
2. гряда (*S* = 13) 6:5 9:20 14:32 15:73 16:7 20:28
3. есть (*S* = 5) 5:5 9:9 10:6 12:11 13:17 16:5 19:17 21:12 25:5 31:43
4. жареный (*S* = 30) 5:10 6:8 9:27 12:12 16:6 19:26 21:16 31:6
5. запах (*S* = 8) 3:5 4:10 6:5 17:46 21:6 23:18
6. зеленый (*S* = 8) 2:5 4:8 5:5 7:5 16:7
7. золоченый (*S* = 12) 6:5
8. изюм (*S* = 11) 19:9

[1] Заметим, что включение слова в ту или иную сферу определяется не местом его в абстрактном «тезаурусе», а именно связями в тексте, ср. *десятина*, *четверть*, *пройти*, *назад*, *душа* (адм.), *миллион*.

9. капуста (*S* = 12) 2:20 3:9 4:27 10:18 12:4 14:54 15:43 16:49 19:19 20:27 21:6 23:9 25:33 31:9
10. квас (*S* = 14) 3:6 9:18 13:8 16:21 17:4 19:10 21:6 23:4 25:12 31:22
11. колчан (*S* = 16) 24:9 26:66
12. кусок (*S* = 6) 3:11 4:12 9:4 16:7 19:31 21:10 25:11 31:119
13. ложка (*S* = 6) 3:17 10:8 19:7 31:15
14. морковь (*S* = 22) 2:32 9:54 15:26 16:11 19:11 209:51 21:6 25:8
15. огород (*S* = 7) 2:73 9:43 14:11 16:12 20:13
16. огурец (*S* = 11) 2:7 3:5 4:6 6:7 9:49 10:21 12:4 14:11 15:12 19:9 20:14 21:9 23:15 25:90 31:18
17. пахнуть (*S* = 4) 1:11 5:46 21:11 23:12
18. печь (*глагол*) (*S* = 7) 19:47 31:45
19. пирог (*S* = 17) 3:17 4:26 8:9 9:19 10:10 12:31 13:7 14:11 16:9 18:47 21:12 31:10
20. редька (*S* = 22) 2:28 9:27 14:51 15:13 16:14
21. рыба (*S* = 7) 3:12 4:16 5:6 9:6 10:6 12:10 14:6 16:9 17:11 19:12 23:8 25:38 31:14
22. саадак (*S* = 48) 27:8
23. свежий (*S* = 4) 5:18 9:9 10:4 16:15 17:12 21:8 25:8
24. седло (*S* = 4) 11:9 26:6
25. соленый (*S* = 8) 3:5 9:33 10:12 14:816:90 21:38 31:9
26. стрела (*S* = 51) 11:66 24:6 29:6 27:5 29:9 30:54
27. стрелять (*S* = 4) 22:8 26:5
28. татарин (*S* = 8) 29:18
29. татарский (*S* = 5) 26:9 28:18
30. тетива (*S* = 48) 26:54
31. хлеб (*S* = 8) 3:43 4:6 9:9 10:22 12:119 13:15 15:3 16:18 18:45 19:10 21:14 25:9

Среди ассоциатов левой части таблицы, в свою очередь, намечаются три подобласти — ЗАПАХ (*запах*, *пахнуть*, *вонять*), ОГОРОД (*огород*, *гряда*, *капуста*, *морковь*, *редька*) и ЕДА (остальные слова). Слово *огурец* соединяет вторую и третью подобласти.

Слово *Телль*, связанное с *лук* ($S = 28$), не показало связей с другими ассоциатами; оно характерным образом связано со словом *стрелок* ($S = 10$), последнее же связано со *стрелять* ($S = 12$). Таким образом, слово *Телль* должно было бы попасть в правую часть таблицы. Ошибкой можно считать положение слова *золоченый*, попавшего в левую часть из-за своей «цветовой» связи с *зеленый* ($S = 5$). С «семантической» точки зрения оно связано с *лук* ‘оружие’ ($S = 12$).

Типы таблиц, перечисленные выше, могут найти себе применение в особых частных случаях, но они не годятся для униформного отражения сотен и тысяч слов, входящих в сеть текстуальных связей. Здесь нужны более простые решения.

В большой табл. 6-А представлено в алфавитном порядке 3000 слов (в электронной версии — 5000 слов и около 300 тысяч текстуальных связей). Покажем структуру словарной статьи на примере слова *стыд*.

Первая строка дает основное слово и три числовых показателя:

$F = 873$ — частоту основного слова,

$A = 109$ — число его текстуальных связей, начиная с $S = 3$,

$[S > 3]$ — минимальный порог, выше которого текстуальные связи включаются в статью.

Далее следуют ассоциаты со значением S их связи с основным словом.

стыд $F = 873$ $A = 109$ $[S > 3]$

ангельский (5), боязнь (8), броситься (4), воспоминание (4), вспыхнуть (6), выступить (4), гнев (4), гордость (6), горе (7), горечь (4), девичий (11), деться (4), досада (4), жгучий (13), жечь (4), забыть (4), закрывать (4), закрыться (4), залить (9), застрелиться (4), злоба (5), испуг (4), испытывать (4), краска (17), лицо (4), ложный (24), малодушие (9), мучительный (4), негодование (8), незаслуженный (5), нерушимый (4), низость (4), нуль (4), оскорбление (5), отчаяние (10), подлый (5), позор (21), позорный (5), покраснеть (5), покрыть (9), поношение (6), потерять (4), поцелуй (4), почувствовать (5), презрение (6), препроводить (5), раскаяние (18), рдеть (5), румянец (4), самолюбие (5), сгорать (16), сгореть (28), сердце (4), слеза (7), совеститься (7), совесть (22), сознаваться (5), сознание (4), спасти (4), срам (21), страх (7), стыдиться (10), стыдно (15), трусить (4), ужас (4), унижение (10), честь (4), чувство (14), чувствовать (5), щека (4)

В поисках интересующего читателя слова следует обратиться к табл. 6-А. Если слово там не найдено, нужно перейти к полной табл. 6-С.

6.4. Частота слова и число его текстуальных связей

Основными параметрами слова в нашем исследовании являются два только что введенных показателя:

F — частота (число адресов) слова и

A — активность слова, т. е. число его текстуальных связей, начиная с $S = 3$.

Смысл нашей основной формулы заставляет предполагать, что оба показателя должны положительно коррелировать друг с другом. В целом эта корреляция выдерживается. У слова *па* $F = 57$, $A = 13$ — со связями *выделывать* (93), *танцевать* (16), *отчетливый* (9), *танец* (8), *ножка* (6), *дама* (5), *зал* (5) и т. д. Рассмотренное выше *волосы* имеет $F = 3156$, $A = 912$. Некоторое общее представление о соотношении двух показателей дает табл. 6.6, где сравниваются два ранговых словаря: один — ранжированный по убыванию частоты, второй — по убыванию активности.

Таблица 6.6

Общее сравнение ранговых словарей частоты и активности

r	F	A	r	F	A
10	20000	1200	2000	670	130
50	10000	830	2500	510	110
100	6700	630	3000	410	97
200	4300	500	4000	290	79
300	3300	420	6000	170	48
400	2700	370	8000	120	33
600	1900	300	10000	80	24
800	1700	250	15000	40	15
1000	1400	220	20000	24	10
1500	900	160			

Это значит, что десятое место в ранговом словаре частоты имеет $F = 20000$, а в ранговом словаре активности — $A = 1200$. Соответственно у трехсотого места $F = 3300$ (в шесть раз меньше) и $A = 420$ (в три раза меньше).

Верхушка ранговых словарей показана в табл. 6.7. Включены слова, относящиеся к первой сотне. В левой части таблицы сосредоточились слова со сравнительно небольшими расхождениями рангов в обоих словарях. В центре даны слова, у которых ранг в ранговом словаре частоты в 9 и более раз превышает ранг в словаре активности. В правой части представлены слова с обратным соотношением рангов.

Таблица 6.7

Ранги слов в ранговых словарях частоты и активности

	rF	rA		rF	rA		rF	rA
рука	4	5	белый	165	1	говорить	1	312
глаз	9	6	лицо	20	2	знать	2	614
голова	11	13	вода	168	4	дело	3	59
жизнь	23	15	черный	185	7	время	5	117
дом	22	25	лошадь	152	9	человек	6	112
вдруг	21	32	берег	413	10	стать	7	303
комната	45	16	окно	119	11	хотеть	8	442
большой	39	28	лес	255	12	видеть	10	1852
сидеть	34	37	воздух	393	18	думать	12	1111
люди	19	53	дерево	544	19	слово	14	261
год	52	24	стена	322	21	спросить	15	200
место	29	50	солнце	512	27	отвечать	16	215
деньги	47	39	широкий	399	29	пойти	17	709
голос	43	44	длинный	342	31	хороший	18	631
иметь	30	64	красный	333	32	сделать	24	591
князь	37	69	небо	473	34	любить	27	346
сторона	40	84	река	578	38	делать	28	1166
город	88	46	густой	714	40	смотреть	33	402
сердце	55	80	яркий	727	42	выйти	38	421
душа	54	85	зеленый	823	44	больше	44	1472
тихий	95	48	ветер	739	49	прийти	48	539
сила	87	68	трава	936	56	много	49	504
новый	51	108	серый	864	57	продолжать	53	637
Бог	42	123	звук	633	63	лучше	57	1311