

Текстуальные связи слов в русской прозе середины XIX в.

Анатолий Я. Шайкевич

*Институт русского языка им. В.В. Виноградова РАН,
Москва, Россия, cfrl@rli.ru*

Аннотация. Корпус русской прозы (14 миллионов словоупотреблений) делится 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 ходе анализа возникает огромная сеть текстуальных связей слов.

Ключевые слова: дистрибутивно-статистический анализ, текстуальные связи слов

Для цитирования: Шайкевич А.Я. Текстуальные связи слов в русской прозе середины XIX в. // Вестник РГГУ. Серия «История. Филология. Культурология. Востоковедение». 2018. № 12 (45). С. 10–31.

Textual links of words in mid-19th century Russian prose

Anatolii Ya. Shaikovich

*V.V. Vinogradov Institute of Russian Language of the Russian Academy of Sciences,
Moscow, Russia, cfrl@rli.ru*

Abstract. A corpus of Russian prose (14 million running words) is divided into fragments of equal size (40 words each). When actual co-occurrence of two words exceeds significantly the expectation, calculated on the basis of the null hypothesis, the two words are said to be linked together. A huge network of textual links of words is constituted as a result of the analysis.

Keywords: distributional statistical analysis, textual links of words

For citation: Shaikovich AYa. Textual links of words in mid-19th century Russian prose. *RSUH / RGGU Bulletin. "History. Philology. Cultural Studies. Oriental Studies" Series.* 2018;12(45):10-31.

Введение

Ставится задача дистрибутивно-статистическими средствами выявить связи слов в корпусе текстов.

Дистрибутивно-статистическим анализом (ДСА) называем формальный (без обращения к смыслу) метод изучения больших собраний текстов, опирающийся исключительно на статистику распределения графических слов (цепочек букв между пробелами). Вводимое понятие **интервала текста** позволяет надеяться на получение самых разнообразных содержательных результатов. Малые интервалы задаются естественным членением текста: комбинаторика букв внутри графических слов дает нам **микроинтервал**, бинарные словосочетания являются объектом анализа в **минимальном интервале**. Эти интервалы приводят к открытию грамматики и лексических единиц, больших чем слово (см. [1]).

* * *

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 S 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. В настоящей статье излагаются результаты, полученные на **среднем интервале**.

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* – *Куницево*, *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:

Таблица 1

Пример машинного представления

волосы густой	195	3156	1687	48	13,90
лежать постель	281	5303	2142	45	29,66
лес куст	99	3692	1022	28	9,85

Первый 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 (величина S округляется до меньшего целого числа), в пятом – величина m . В ходе работы компьютера возникает грандиозная таблица, которая включает более 36 тысяч слов и примерно 640 тысяч текстуальных связей.

Даже двукратное совместное появление пары слов может свидетельствовать о вполне реальных семантических связях. Слово *жмурки*, например, по два раза встретилось со словами *вздумать*, *давай*, *колокольчик*, *побежать*, *прятаться*, *разобрать*, *ребята*, *сзади*; по три общих адреса имеет оно со словами *бегать*, *глаз*, *завязанный*, *игра*, *играть*, *ловить*, *товарищ*, *угол*. Все эти текстуальные связи показывают S не менее 3.

Для восприятия человеком компьютерная выдача преобразуется в более простой вид. Покажем структуру словарной статьи на примере слова *стыд*. Первая строка дает основное слово и три числовых показателя: $F = 873$ – частоту основного слова; $A = 109$ – его «активность», т. е. число текстуальных связей. Начиная с $S = 3$ [$S \geq 3$] – минимальный порог, выше которого текстуальные связи включаются в статью. Далее следуют **ассоциаты** со значением S их связи с основным словом.

Пример 1. *Стыд* $F = 873$ $A = 109$ [$S > 3$]

ангельский (5), *боязнь* (8), *броситься* (4), *вспоминание* (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), *закраться* (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).

Средний интервал во многом наследует результаты минимального интервала, ср. пару *волосы* – *густой*, пары *ахиллесова пята* ($S = 170$), *выведенное яйцо* ($S = 129$), *несолоно хле-*

бать ($S = 123$), *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*¹. Однако большинство найденных пар обнаруживаются 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*, *немый*, *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*, *всклоченный*, *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 проявляется весьма характерная для прозы совокупность слов, описывающих внешность человека. Такая совокупность (**кластер**) скорее исключение в прозе. К идее кластера мы еще вернемся ниже.

Эмпирически полученная сеть связей представляет собой яркий контраст с умозрительно построенными тезаурусами, обычно предполагающими дискретное устройство словаря. Тем не менее реальная семантическая дискретность может проглядывать и в сети текстуальных связей. Примером могут служить слова *магазин* и *лавка*. У первого слова зафиксировано 198 текстуальных связей, у второго слова находим 260 связей. Оба слова непосредственно связаны друг с другом ($S = 11$), кроме того есть 38 слов, одновременно связанных с двумя нашими словами (связи второго порядка): *базар*, *бакалейный*, *бульвар*, *вывеска*, *город*, *гостинный*, *европейский*, *зайти*, *китаец*, *китайский*, *книжный*, *колонияльный*, *купец*, *купить*, *лавочка*, *мастерская*, *материя*, *мостовая*, *окно*, *переулок*, *покупатель*, *покупать*, *приказчик*, *прилавок*, *продаваться*, *продавец*, *ре-*

¹ 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.

сторан, табачный, товар, торговать, торговец, торговля, торговый, тротуар, узенький, улица, фруктовый, харчевня.

Казалось бы, семантическая близость двух слов налицо. Однако обратившись к текстуальным связям слова *лавка* и на минуту отказавшись от «асемантического» формализма, мы тут же обнаружим аномалию. В списке 16 связей, упорядоченных по уменьшению *S*, находим связи со словами, никак не вписывающимися в сферу торговли (ниже они отмечены звездочкой): *мелочной* (*S* = 51), *товар, книжный, изба*, полати*, овощной, печь*, покупатель, бакалейный, гостинный, мясной, купец, приказчик, сидеть*, торговать, пол**. Идя дальше по списку, найдем и другие аномальные слова: *присесть, стряпной, подостлать, зыбка, печка, садиться, сесть*. Знание русского языка подскажет, что *лавка* соотносится с двумя семантическими объектами: 1) торговая точка, 2) скамья. Рассмотренная аномалия есть лишь отражение этого факта в сети текстуальных связей.

В процессе настоящего исследования для нескольких десятков слов мы отошли от строгого формализма и с учетом нашего знания языка создали «искусственные» слова, соответствующим образом маркируя их в тексте. В данном случае текстуальные связи определены для двух слов *лавка-1* (*F* = 1020, *A* = 293) и *лавка-2* (*F* = 470, *A* = 165). Первое слово сохраняет «торговые» связи, второе – значительно увеличивает *S*, расширяет круг «бытовых» связей (любопытны связи со словами *Кутузов* и *Пугачев*).

По текстуальным связям нетрудно понять мнемонику следующих «искусственных» слов: ***коса-ж*** (*A* = 213): *лента, заплетать, вплести, заплести, расплести, заплетенный, русский, прядь, растрепаться, кокошник, расчесывать, густой, обрезать, длинный, волосы, монисто* и др.; ***коса-сх*** (*A* = 60): *точило, косить, серп, косец, лезвие, точить, брусок, грабли, косарь, цеп, трава, топор, кузнец* и др.; ***коса-в*** (*A* = 6): *песчаный, таможенный, неводчик, камыш, мелькать, море*.

Создание «искусственных» слов может быть мотивировано с учетом дискурсивных функций, ср.: ***прощать*** (*A* = 54): *простить, прощение, извинять, вина, виноват, враг, грешник, Бог, великодушный, забывать* и др.; ***прощай*** (*A* = 242): *уходить, увидеться, свидание, лихо, целовать, завтра, провожать, целоваться, спасибо, заболтаться, кланяться, пора, пойти, поцеловать, проститься, уйти* и др. Можно полагать, что наше исследование остается на 90% формальным даже с учетом этих «семантических» отступлений.

Смысл нашей основной формулы заставляет предполагать, что оба показателя *F* и *A* должны положительно коррелировать

друг с другом. В целом эта корреляция выдерживается. У слова *па* ($F = 57$) обнаружено 13 текстуальных связей: *выделявать* ($S = 93$), *танцевать* ($S = 16$), *отчетливый* ($S = 9$), *танец* ($S = 8$), *ножка* ($S = 6$), *дама*, *зал* ($S = 5$), *лучший* ($S = 4$), *девица*, *довольно*, *мелкий*, *непреренно*, *учитель* ($S = 3$). У частых слов и A велико. Вот несколько характерных примеров:

Таблица 2

Корреляция показателей F и A

слово	F	A	слово	F	A
белый	4894	1457	лицо	15946	1443
вода	5105	1310	лошадь	5682	1233
волосы	3156	912	нога	8319	1238
дверь	8475	1046	русский	6700	1005
дорога	6066	913	солнце	2293	997
лес	3692	1150	стоять	13716	1358

Для выявления аномалий в этой основной тенденции удобно сравнивать не F и A , а ранг r в словаре, ранжированном по частоте (rF), и ранг в словаре, ранжированном по активности (rA). В левой части таблицы 3 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 (слова с повышенной активностью). В правой части представлены слова с обратным соотношением рангов (слова с пониженной активностью).

Таблица 3

Ранги слов в ранговых словарях частоты и активности

слово	rF	rA	слово	rF	rA	слово	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

Окончание табл. 3

слово	rF	rA	слово	rF	rA	слово	rF	rA
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
час	56	113	рыба	807	73	заметить	64	838
мысль	71	101	снег	956	77	понимать	65	877
ночь	93	81	цвет	1099	93	кажется	69	1605
более	50	138	высокий	198	22	простой	70	872
лежать	138	51	волосы	329	36	остаться	75	985

Поражает преобладание глаголов в правом столбце (в левом столбце их всего три, в центральном – ни одного). Причина, по-видимому, кроется в синтаксической роли глаголов. При общем богатстве актантов у глаголов самыми важными оказываются позиции субъекта. Чаще всего в этой роли выступают обозначения человека, выражаемые либо антропонимами (слабо представленными в материале данного исследования), либо местоимениями (совсем в нем не представленные). В центральном столбце поразительно преобладание существительных, указывающих на природу. Отметим также прилагательные – обозначения цвета.

При группировке связей слов *лавка* и *коса* мы опирались на семантические соображения. Однако при анализе сети текстуальных связей можно испробовать и более формальные приемы. В таблице 4 пронумерованы ассоциаты слова *лук* и показаны их связи друг с другом. Поясним структуру таблицы на примере слова *запах*:

Пример 2

запах ($S = 8$) 3:5 4:10 6:5 17:46 21:6 23:18

Значимость его связи с основным словом *лук* указана в скобках – в данном случае ($S = 8$); далее следуют номера слов, с которыми связано слово *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 4

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	
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	

Окончание табл. 4

14. морковь ($S = 22$) 2:32 9:54 15:26 16:11 19:11 20:9 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	22. саадак ($S = 48$) 27:8
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	24. седло ($S = 4$) 11:9 26:6
20. редька ($S = 22$) 2:28 9:27 14:51 15:13 16:14	26. стрела ($S = 51$) 11:66 24:6 29:6 27:5 29:9 30:54
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	27. стрелять ($S = 4$) 22:8 26:5
23. свежий ($S = 4$) 5:18 9:9 10:4 16:15 17:12 21:8 25:8	28. татарин ($S = 8$) 29:18
25. соленый ($S = 8$) 3:5 9:33 10:12 14:8 16:90 21:38 31:9	29. татарский ($S = 5$) 26:9 28:18
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	30. тетива ($S = 48$) 26:54

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.

В добавление к таблице 4 упомянем еще слово *Тель*, связанное с *лук* ($S = 28$), которое 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 *Тель* 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’ ($S = 12$).

Слово *лук* самым естественным образом распалось на два омонима. По-другому обстоит дело со словом *бедный*. Среди ассоциатов этого слова (с $S > 5$) ярко выступают две пары, которые могут считаться своеобразными полюсами небольшого семантического пространства, пара *бедность* – *богатство* и пара *слеза* – *плакать*. Одна группа ассоциатов своими перекрестными связями показывает, что тяготеет к первой паре, – это *бедняк, богатый, богач, дворя-*

нин, дворянка, деньги, жениться, жить, мелкопоместный, милостыня, небогатый, нищета, нищий, пожалеть, раздать, родственница, семейство, семья. Ко второй паре тяготеют *больной, дитя, жалкий, жаль, зачахнуть, несчастный, ребенок, умереть, чахотка.* В промежуточном поле оказываются *беспомощный, благородный, девочка, девушка, дети, добрый, женщина, люди, мальчик, мать, одетый, помогать, помочь, сирота, сиротка, старушка, человек.* С абстрактной точки зрения слово *вдова* может казаться в равной мере привязанной как к «социальному», так и к «меланхолическому» полюсу, однако некоторые его связи (*богатый, дворянка, жениться, жить, небогатый, родственница*) показывают, что в нашем корпусе *вдова* появляется преимущественно в матримониальных контекстах.

Тщетной была бы попытка просчитать подобным образом наш материал; лишь для частых слов такая задача кажется осуществимой. Ниже будут приведены некоторые примеры группировки текстуальных связей. Значения *S* приводятся только при $S > 9$, а в квадратных скобках дается самый значимый ассоциат основной группы связей. Вот самый простой случай – уникальная связь одного из омонимов:

Пример 3

брак – браковать (13) [законный]; *Десна – купальня* (20) [зуб];
устье – печь (11) [Ока]; *почка – легкое* (46) [древесный].
гитара – извозчиций (10) [струна];

Менее частый из двух омонимов может иметь и несколько связей:

Пример 4

ласка – горностай (39), *зверек* (15), *хищный* [нежный];
болтун – утенок (20), *яйцо* [Герцен].

Такой способ представления может применяться и к случаям полисемии:

Пример 5

поезд – дружка (34), *поезжане* (23), *свадебный* (21), *молодые* (14), *процессия, сваха, венчание* [вагон];
найтись – ответить, молвить, растеряться, возразить, согласиться [поискать];
холостой – заряд (29), *гнездо, зарядить, стрельба, выстрелить* [женатый];

гряда – бурун (42), тянуться (22), риф (17), свинцовый (11), рейд (11) [огород];

звезда – лента (24), фрак (23), камергерский (13), сановный (13), анненский (12), бриллиантовый (12), крест (12), регалия (11), Станислав (11), мундир (10), орден (10) [небо];

кислый – гримаса (32), мина (25), скорчить (21), физиономия (11), улыбка, сморщенный, постный, неудовольствие, рожа [капуста];

скосить – презрительный (11), глаз (10), вопросительный, губа, презрение [луг];

ссылка – цитата (12) [тюрьма];

ствол – ружейный (52), ружье (38), порох (18), стрелять [дерево];

краска – залить (28), лицо (23), выступить (19), стыд (17), проступить (11), щека (11) [выкрашенный];

грубый – сукно, ткань, пестрядинный, волокно, изделие [дерзкий];

двор – преображенский, принцесса, фрейлина, принц [ворота];

выходить – проверка, zero, значит, ан, гадать, жених [замуж].

Далеко не всегда ясно, какую из групп связей следует считать основной. В таких случаях можно прибегнуть и к менее экономному способу представления.

Пример 6

ветреный $F = 137$ $A = 23$

1. *погода (19), дождливый (11), сырой, холодный, морозный, дождь;*

2. *легкомысленный (17), пустой, любить, мальчишка, увлекаться, женщина.*

вкус $F = 1136$ $A = 276$

1. *мясо (18), дичина (16), блюдо, обоняние, пища, приторный, ягода, гастроном, запах, кушанье, лафит;*

2. *приятись (25), эстетический (25), убранный (24), изящный (22), меблировать (17), изящество (16), одетый (15), стиль (13), неприхотливый (11), одеваться (10), рококо (10), туалет (10).*

Заметим, что в перечисленных выше случаях не ставилась цель решать традиционные вопросы «семантической» лингвистики: полисемия или омонимия, одно значение или два. Велся лишь поиск группировки связей, полученных из текста.

Сходный подход возможен не только к отдельному слову, но и к группам слов. Начнем с синонимов. Исходя из общих соображений, мы ожидаем, что синонимы должны обладать каким-то набором общих текстуальных связей. Число общих связей зависит от

активности синонимов (A), прежде всего – от активности **менее частого** синонима. Слово *Англия* ($A = 196$) и слово *Великобритания* ($A = 4$) связаны друг с другом крайне значимой текстуальной связью ($S = 23$). Ясно, что число общих связей с третьими словами у этой пары не может быть больше четырех. В реальности обнаружено одно слово *Россия*, связанное и с *Англия* ($S = 16$), и с *Великобритания* ($S = 4$). Только одну общую связь находим в паре *Амур – Купидон* ($S = 15$). Это слово *стрела*, связанное и с *Амур* ($S = 16$), и с *Купидон* ($S = 10$).

В общем случае синонимы могут и не иметь непосредственной связи друг с другом. Так, *базар* ($F = 349$, $A = 115$) и *рынок* ($F = 241$, $A = 80$) не связаны друг с другом, но имеют 15 общих ассоциатов: *город, купить, лавка, лавочка, площадь, покупать, провизия, продавать, продажа, товар, торг, торговать, торговец, трактир, улица*.

Итак, наши синонимы *базар* и *рынок* имеют пятнадцать общих связей ($L = 15$). Много это или мало? Удобным общим знаменателем Q будет служить корень из числа связей менее частого синонима: $Q = \sqrt{A_{\min}}$. Соотношение $K = L/Q$ и окажется показателем тесноты связей синонимов. В нашем случае для слов *базар* и *рынок* $K = 1,7$. Сравним этот результат со следующими парами синонимов:

Пример 7

блестеть ($A = 235$) и *блистать* ($A = 130$)

белизна, белый, бледный, блеск, бриллиант, глаз и т. п.

$L = 42$, $Q = 11,4$, $K = 3,7$.

внезапный ($A = 229$) и *неожиданный* ($A = 108$)

быстро, вдруг, вздрогнуть, впечатление, вскочить и т. п.

$L = 37$, $Q = 10,4$, $K = 3,6$.

вокруг ($A = 432$) и *кругом* ($A = 433$)

белый, береза, сверху, вдали, везде, веять, внизу, волна и т. п.

$L = 83$, $Q = 20,8$, $K = 4,0$.

велеть ($A = 407$) и *приказать* ($A = 339$)

барин, барыня, батюшка, бутылка, везти, вино, водка и т. п.

$L = 106$, $Q = 18,4$, $K = 5,8$.

Текстуальные связи двух последних синонимов отражают существенную социальную сторону изучаемого корпуса – Россию середины XIX века. Здесь и субъект волеизъявления (*барин, барыня, батюшка, государь, граф, губернатор*), и фигуры подчиненных (*горничная, камердинер, кучер, лакей, повар, приказчик, слуга*,

управляющий), и громадное число глаголов, обозначающих желаемое действие (*выслать, доложить, зажарить, запереть, запечатать, наказать, написать, обождать, поставить, почистить* и др.). Бросается в глаза и связь с мегакластерами, о которых пойдет речь ниже: ЛОШАДЬ (*ехать, заложить, запрячь, карета* и т. п.) и ЗАСТОЛЬЕ (*завтрак, кушать, самовар* и т. п.).

Хрестоматийная пара синонимов *конь* ($F = 1388$, $A = 551$) и *лошадь* ($F = 5682$, $A = 1233$) связаны друг с другом непосредственно ($S = 32$) и бьют все рекорды по числу общих связей и по тесноте взаимной синонимичности: $L = 211$, $Q = 23,5$, $K = 9,0$.

У большинства ассоциатов их связи с парой основных синонимов *конь* – *лошадь* не слишком различаются по величине S (здесь первой дается S связи с *конь*, второй – связи с *лошадь*): *барышник* 11/10, *буланый* 9/10, *верста* 7/20, *верхом* 16/33, *гнедой* 11/19, *гусар* 7/9, *казак* 18/20, *конюшня* 17/30, *копыто* 26/42, *повод* 39/67, *сено* 11/21, *слезать* 16/16, *топот* 13/24, *тройка* 19/38, *узды* 21/33, *фыркать* 20/20, *хвост* 9/14. В то же время десятки ассоциатов тяготеют к слову *конь* но не *лошадь*: *амазонка* 13/6, *аргамак* 19/6, *атаман* 6/–, *битва* 7/–, *боевой* 19/–, *боярин* 9/–, *боярыня* 21/–, *броня* 7/–, *взвиться* 20/11, *взмывленный* 21/8, *витязь* 16/–, *вороной* 36/30, *всадник* 61/21, *грива* 29/22, *доброезжий* 13/6, *драбант* 12/–, *дыбы* 22/17, *заржать* 20/17, *калмык* 7/4, *кобылица* 15/–, *конюх* 26/10, *красивый* 7/4, *круп* 8/5, *кудесник* 11/–, *лихой* 16/6, *лука* 26/–, *молодец* 12/4, *мчаться* 19/15, *опор* 9/7, *опричник* 15/–, *пеший* 17/6, *пика* 6/–, *погоня* 19/6, *подскакать* 14/13, *пришпорить* 25/10, *Пугачев* 20/12, *рейтор* 6/–, *ретивый* 6/–, *сабля* 17/7, *седло* 43/34, *сокольник* 6/–, *спешиться* 20/11, *станция* 6/–, *степь* 16/6, *стрела* 15/–, *стреля* 13/11, *тавро* 10/–, *татарин* 8/4, *удила* 21/5, *узды* 19/11, *улан* 14/10, *форейтор* 11/10, *холоп* 13/–, *хорунжий* 7/4, *царевич* 13/– и т. п. Все эти слова предполагают верховую езду, часто видны стилистическая приподнятость и исторические коннотации.

У слов, скорее тяготеющих к *лошадь*, счет идет на сотни; далее следует очень небольшой список примеров: *барин* –/10, *велеть* –/20, *возжечь* 6/28, *воз* –/25, *двор* 7/24, *денник* –/16, *дорога* 7/36, *дрожки* –/19, *ехать* 11/37, *закладывать* –/29, *запрягать* 7/47, *кавалерист* –/10, *карета* 7/22, *кибитка* 9/20, *кнут* 4/24, *кобыла* –/10, *козлы* 4/24, *колесо* 8/24, *корова* –/32, *крыльцо* 7/21, *купить* –/14, *кучер* 8/61, *ломовой* –/12, *мужик* –/19, *но(!)* –/12, *облучок* 4/15, *обоз* –/22, *оброть* –/10, *обывательский* –/21, *овес* 7/28, *охотник* –/8, *повозка* –/27, *полоз* –/16, *почтовый* –/23, *пошевни* –/9, *править* 5/22, *присяжная* –/24, *рысь* 9/40, *сани* 6/35, *сарай* –/12, *снег* –/16, *собака* 4/13, *станция* 4/24, *стойло* 4/13, *таран-*

тас –/20 *телега* 4/43, *хомут* –/16, *шля* –/11, *экипаж* 7/38, *ящик* 12/44, *ясли* –/11. В этом перечне на первый план выходят факты реальной действительности, иногда видна и стилистическая приземленность (*брюхо, закладывать, кляча, лошаденка, одер, мериннок, плестись, тащиться*).

В большой сети парных текстуальных связей неизбежно должны встречаться какие-то множества n слов, у которых число связей (Lk) превышает $(n - 1)$. Иначе говоря, в соответствующей графе наблюдаются циклы. В той мере, в какой Lk приближается к максимальному $n(n - 1)/2$, можно говорить о сгустке или **кластере** связей. Как пример рассмотрим матрицу текстуальных связей пяти слов с их значениями S .

Таблица 5

Матрица текстуальных связей слов *год, весна, лето, осень, зима*

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 этих слов. 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* ($S = 12$).

Можно полагать, что сами кластеры связей обязаны своим существованием разным факторам – как общеязыковым, так и текстовым (в частности – сюжетным).

Снова обратимся к текстуальным связям слова *дуэль*, где можно ожидать сравнительно компактный круг ассоциатов, привязанных к этому фрагменту внеязыковой действительности. Опираясь на связи с высокими S , сформируем сначала **центр кластера**, объем которого допускал бы построение матрицы, удобной для печатного издания. В данном случае центр включает десять слов (см. табл. 6). К этим десяти словам присоединим 35 слов, имеющих хотя бы две связи с центром (первое пополнение), например, как это сделано в табл. 7.

Таблица 6

Центр кластера ДУЭЛЬ – ПОЕДИНОК

	вызов	выстрел	драться	дуэль	пистолет	поединок	противник	ранить	секундант	убить
вызов	–	–	–	44	–	10	3	–	16	4
выстрел	–	–	–	5	29	–	3	36	–	14
драться	–	–	–	37	5	11	10	3	21	7
дуэль	44	5	37	–	28	8	20	10	56	20
пистолет	–	29	5	28	–	–	14	6	27	20
поединок	10	–	11	8	–	–	20	–	34	5
противник	3	3	10	20	14	20	–	3	28	–
ранить	–	36	3	10	6	–	3	–	3	36
секундант	16	–	21	56	27	34	28	3	–	–
убить	4	14	7	20	20	5	–	36	–	–

Таблица 7

Первое пополнение кластера ДУЭЛЬ – ПОЕДИНОК

	вызов	выстрел	драться	дуэль	пистолет	поединок	противник	ранить	секундант	убить
барьер	–	9	–	–	38	–	17	–	21	–
Кавказ	–	–	–	4	–	–	–	4	–	5
курок	–	15	–	–	32	–	–	–	–	12
отмерить	–	–	–	–	14	–	8	–	–	–
офицер	–	8	6	–	5	7	–	25	6	5
пуля	–	28	3	4	28	–	–	31	5	10
стрелять	–	–	4	6	34	–	4	14	6	13
стреляться	–	–	–	10	24	–	–	–	25	–
шага	–	–	11	3	3	36	29	9	–	–

Слова первого пополнения, в свою очередь, могут быть связаны друг с другом: *барьер* – с *отмерить*, *пуля* – с *курок*, *стрелять*, *стреляться*, *шага*. Во втором пополнении окажутся слова, связанные с первым пополнением, и слова, связанные с центром одиночными связями, – еще 13 слов, например *повод*, *скандал*, *шаг*.

Таким образом, в кластере оказалось 58 слов. Число текстуальных связей слова внутри кластера (a) снова сопоставим с $\sqrt{A}$. Семнадцать слов, у которых (a) $> \sqrt{A}$, могут быть названы **ядром кластера**. С интуитивной точки зрения к ядру хотелось бы присоединить слова *шага*, чье соотношение (a) и ($\sqrt{A}$) очень близко к выбранному порогу.

Таблица 8

Ядро кластера ДУЭЛЬ – ПОЕДИНОК

	A	a		A	a
барьер	24	7	поединок	36	13
вызов	42	11	прицельиться	8	7
выстрелить	91	13	противник	105	17
драться	97	14	пуля	179	15
дуэль	81	34	ранить	286	18
застрелить	38	8	секундант	45	18
курок	35	8	стреляться	15	4
отмерить	11	4	убить	392	23
пистолет	135	25			

Общий итог не должен затемнять некоторые важные подробности. При общности дуэли с другими формами вооруженного конфликта (*выстрел*) ядро кластера помимо двух главных слов (*дуэль* и *поединок*) включает технические детали (*барьер*, *отмерить*, *секундант*) и ряд чисто лингвистических особенностей – *драться*, *стреляться*. Краткосрочность дуэли объясняет склонность к совершенному виду *выстрелить*, *убить* vs *стрелять*, *убивать*. Любовь нарратива к деталям объясняет присутствие слов *пуля*, *курок*. Обратим внимание на связь *поединка* со *шагой* (и с *фехтованием*), отсылающую нас к XVIII веку. Любопытна концентрация слов причины и следствий дуэли во втором пополнении (*клевета*, *обида*, *повод*, *ссора*, *столкновение*, *история*, *кончиться*, *разжаловать*, *скандал*).

По сравнению с кластером дуэли максимально компактным и автономным кажется кластер ДОКТОР – БОЛЕЗНЬ. К центру из

семи слов (все они входят в ядро кластера) добавляются 140 слов в первом пополнении. С учетом второго пополнения общий объем кластер достигает 250 слов, причем большинство их относится к ядру.

Таблица 9

Центр кластера ДОКТОР – БОЛЕЗНЬ

	доктор	врач	лечить	лекарство	больной	болезнь	умереть
доктор	—	23	42	28	43	30	8
врач	23	—	26	7	32	14	4
лечить	42	26	—	35	18	34	3
лекарство	28	7	35	—	29	23	—
больной	43	32	18	29	—	28	8
болезнь	30	14	34	23	28	—	10
умереть	8	4	3	—	8	10	—

Рассмотренные кластеры несомненно соответствуют группам слов, объективно существующим в русском языке. В нашем корпусе обнаружены и такие локальные кластеры, которые связаны с ролью тех или иных слов в прозе XIX века. Характерным примером может служить круг ассоциатов словосочетания *Невский проспект*.

В центр кластера попали *Невский проспект, гулять, магазин, набережная, рысак, тротуар, экипаж*. Среди 29 слов первого пополнения отметим *великолепный, извозчик, изящный, камелия, карета, кататься, коляска, мелькать, модный, мостовая, мчаться, отель, пешеход, роскошный, толпа, щегольской*. Во втором пополнении находим *бульвар, дрожки, окно, освещенный, подъезд, прохожий*. Атмосфера гуляющих прохожих, модных магазинов, освещенных окон, мчащихся экипажей так и встает за перечисленными связями².

Своеобразным антиподом Невского проспекта выступает Сенная площадь с ее выразительными текстуальными связями: *гауптвахта (22), часть (21), притон (15), окраина (10), разврат (10), заведение (9), нищий (7), голод (6), тюрьма (6)*. Колорит «Петербургских трущоб» Крестовского виден в следующем миникластере.

² Некоторым московским подобием Невского проспекта кажется Кузнецкий мост с его ассоциатами: *бал, бульвар, готовый, купить, магазин, Москва, француженка, экипаж*.

Таблица 10

Центр кластера «Сенная площадь»

	Сенная площадь	Вяземская лавра	обитатель	трущоба	трущобный
Сенная площадь	–	31	17	36	–
Вяземская лавра	31	–	21	–	–
обитатель	17	21	–	25	18
трущоба	36	–	25	–	40
трущобный	–	–	18	40	–

На примере кластеров ДУЭЛЬ – ПОЕДИНОК и ДОКТОР – БОЛЕЗНЬ был опробован путь формирования автономных групп слов в сети текстуальных связей. Принимая текстуальные связи с их значениями S как данное, мы создаем пробный центр – несколько слов, удобных для матричного представления. Наличие перекрестных связей в матрице говорит о перспективности центра как старта в построении какого-то кластера. Далее к этому центру присоединяется пополнение – новые слова, имеющие по несколько связей со словами центра. Затем процедура может повторяться в отношении уже полученной большей группы слов. На этом не слишком формальном пути мы неизбежно столкнемся с проблемой отношений между создаваемыми кластерами.

В качестве примера обратимся к глаголу *пить* – очень частому и активному слову ($F = 3953$, $A = 315$). Вот как выстраиваются его ассоциаты в порядке уменьшения S :

Таблица 11

Ассоциаты глагола *пить*

	S	F	A		S	F	A
чай	129	3756	512	рюмка	34	846	210
вино	77	2243	462	налить	28	437	127
водка	72	1820	410	бутылка	27	1015	300
есть	70	2106	358	чашка	26	941	329
выпить	55	2596	408	вприкуску	24	13	8
кофе	54	699	191	здоровье	24	1726	216
стакан	46	1348	311	вода	20	5105	1310
шампанское	41	774	200	коньяк	20	116	31
наливать	35	359	89	мед	20	307	139
пиво	34	401	129				

Ясно, что среди этих 19 слов есть и такие (с меньшей активностью), у которых мало шансов привлечь пополнение (*наливать, вприкуску, мед, коньяк*). Отложив на время ассоциат *есть*, мы никак не можем игнорировать самое активное слово – *вода*. В качестве пробы рассмотрим центр кластера *пить* в таком составе: *пить, вино, вода, водка, выпить, кофе, налить, пиво, рюмка, стакан, чай*. В соответствующей матрице у слова *вода* оказалось пять лакун (из 10 возможных), среднее значение *S* у него 10 (в целом по матрице оно равно 36); это самый сомнительный кандидат в центр. Чуть больше среднее значение *S* у слова *кофе*.

Оставим в центре восемь слов и в угоду семантике присвоим ему новый ярлык – *выпить*. Центр этого кластера сопровождается мощным первым пополнением (344 слова). Следующие члены первого пополнения связаны со всеми восемью словами центра: *бутылка, залпом, стаканчик, стол, херес, чай, шампанское*. Как видим, слово *чай* упорно восстанавливает свое место в кластере. Двадцать пять слов пополнения связаны с семью словами центра: *винцо, выкушать, выпивать, гость, графин, закуска, капля, квас, кушать, мадера, наливать, напиток, пирог, подать, поднести, поднос, потчевать, потчевать, принести, пьяный, ром, стаканчик, угощать, чашка, штоф*. Слова *есть, закусить, обед, сахар, тарелка* имеют по шесть связей с центром. На наших глазах начинает вырисовываться **мегакластер ЗАСТОЛЬЯ**.

Тема застолья, усердно эксплуатируемая прозой, не единственный источник появления мегакластеров. Тройка слов *дорога, ехать, лошадь* указывает на важную тему европейской прозы до XX века, но детальная лексическая разработка *лошади* – характерная черта именно русской прозы. Вот как выглядит центр мегакластера ЕЗДЫ – *верста, всадник, дорога, ехать, конь, кучер, лошадь, поехать, рысь, сани, скакать, тройка, ямщик*.

Можно предполагать, что в сети текстуальных связей существует еще несколько мегакластеров, например – КОМНАТЫ (*вист, горница, дверь, зеркало, картина, комната, кровать, мебель, обои, образ, окно, пол, портрет, стекло, стена, стол, угол* и т. п.), ДОМА (*ворота, двор, деревянный, дом, забор, изба, калитка, каменный, крыльцо, крыша, сарай, улица...*), ВНЕШНОСТИ (*белокурый, борода, бровь, волосы, высокий, глаз, голова, губа, коса, лицо, морщина, нос, рост, русский, рыжий, седой, смуглый, стройный, усы, фигура, худой, черта...*).

Заключение

Работа ДСА в среднем интервале привела к созданию сети текстуальных связей – новому объекту корпусной лингвистики, открывающему широкие возможности изучения текстов, далеко выходящие за пределы традиционных интересов лингвистов. От изучения корпуса как целого можно переходить к анализу авторских, жанровых и тематических подкорпусов (желательно большого объема).

Слово *лес* у русских писателей встречается с разной частотой. На 100 000 словоупотреблений частота этого слова составляет у Тургенева 24, у Достоевского 6, у Толстого 31. Тем не менее у этих авторов появляется одна и та же тройка слов, взаимно связанных текстуальными связями: *лес, дерево, куст*. У трех авторов слово *лес* связано со словами *кругом* и *поле*. Но за пределами этой пятерки начинаются различия. У Достоевского с двумя другими авторами находим по одной общей связи, Тургенев и Толстой для слова *лес* дают 24 общих связей: *береза, ветер, виднеться, густой, дорога, дуб, зарости, зеленый, земля, лист, небо, овраг, опушка, покрытый, поляна, река, ружье, солнце, сосновый, темнеть, трава, туман, широкий, яркий*.

В общем корпусе прозы слово *виднеться* имеет 580 связей. Лишь одна связь (со словом *черный*) обнаружена в подкорпусах всех трех авторов, у Тургенева и Толстого находим 20 общих связей данного глагола: *белый, ветер, высокий, из-за, из-под, красный, крыша, куст, лес, направо, небо, острый, поле, полоса, сквозь, сплошной, стена, темнота, туман, тянуться*. Фигура «наблюдателя» очень ярко отражена в этой группе слов.

Поскольку ДСА не привязан к конкретному языку, видны перспективы межъязыкового сравнения корпусов текстов. В публикации [2] средний интервал ДСА проверялся на статистически ограниченных корпусах текстов (по 200 000 словоупотреблений). Помимо собственно семантического компонента, характерным элементом текстуальных связей там оказывалась рифма: *человек – век, роза – мороз, очи – ночь, Киприда – обида, Венера – пещера*.

Литература

1. Шайкевич А.Я., Андрущенко В.М., Ребецкая Н.А. Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг.: В 2 т. М.: ЯСК, 2013.
2. Шайкевич А.Я. Сеть семантических текстуальных связей в поэзии Пушкина и Мицкевича // Славянское языкознание. XIII Международный съезд славистов: Доклады российской делегации (Люблина, 2003 г.). М.: Индрик, 2003. С. 576–588.

References

1. Shaikevich AYа. Andrjuscenko VM., Rebetskaya NA. Distributional statistical analysis of the language of Russian prose 1850–1870. In 2 vols. Moscow: YaSK Publ.; 2013. (In Russ.)
2. Shaikevich AYа. Network of semantic textual links in the poetry of Pushkin and Mickiewicz // Slavic Studies. XIII Congress of Slavic Studies. Reports of the Russian delegation (Ljubljana, 2003). Moscow: Indrik Publ.; 2003. p. 576-88. (In Russ.)

Информация об авторе

Анатолий Я. Шайкевич, Институт русского языка им. В.В. Виноградова Российской академии наук, Москва, Россия; 119019, Россия, Москва, ул. Волхонка, д. 18/2; cfrl@rli.ru

Information about the author

Anatolii Ya. Shaikevich, V.V. Vinogradov Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia; bld. 18/2, Volkhonka st., Moscow, Russia, 119019; cfrl@rli.ru