ЧАСТОТНЫЕ ИЗМЕНЕНИЯ ЛЕКСИКИ (НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА)

М.В. Беляков

В статье рассматриваются изменения частотных характеристик лексики русского языка по результатам сравнительного анализа данных, представленных в новом частотном словаре русского языка, и данных, полученных средствами компьютерного лингвостатистического анализа. Выявляются особенности частотных изменений употребления лексики в разных функциональных стилях русского языка.

Ключевые слова: лингвостатистика, частотный словарь, функциональный стиль, компьютерная обработка текста.

FREQUENCY CHANGES IN VOCABULARY (BASED ON THE RUSSIAN LANGUAGE MATERIAL)

M.V. Belyakov

The article deals with some frequency characteristics changes in Russian lexis based on the results of comparative analysis of the data presented in the new frequency dictionary of the Russian language and the data obtained by means of computer linguostatistical analysis. The features of frequency changes in the use of vocabulary in different functional styles of the Russian language are revealed.

Key words: linguistic statistics, a frequency dictionary, functional style, computer text processing.

Введение

Лексика, бесспорно, является наиболее подвижной языковой составляющей, подвергающейся постоянным изменениям, вызванным переменами как лингвистической, так и экстралингвистической ситуации. Существует даже такое понятие, как хронологический шаг, под которым понимаются временные периоды, за которые происходят заметные изменения в лексическом составе языка. Выделяют ускоренный шаг (10–20 лет), умеренный с плавными сдвигами (30–40 лет) и замедленный (50 лет и более). Однако с начала XXI в. наблюдается очевидное ускорение перемен в разных областях жизни социума, которые естественным образом влияют на изменения, происходящие в лексическом пласте языка, причем в лексике, характерной для разных языковых стилей.

1. Количественный анализ лексики

Одним из возможных способов проследить эти изменения является количественный способ определения изменений в частотности употребления лексем за определенный период времени. Целью нашего исследования было выявление этих изменений наиболее объективным количественным методом. Задача – провести эксперимент по установлению количественных изменений в частотности употреб-

ления произвольно выбранных лексем разных частей речи в русском языке за последнее десятилетие при помощи обработки текстов, представленных в Национальном корпусе русского языка.

Провести подобное исследование можно путем обработки текстового материала компьютерными программами автоматизированной обработки текста, включая программы SEO-анализа (Search Engine Optimization). Полученные по результатам обработки количественные данные сравниваются с числовыми показателями, представленными в хорошо известном «Новом частотном словаре русской лексики» (на материалах Национального корпуса русского языка) О.Н. Ляшевской, С.А. Шарова [2]. Словарь вышел в свет в 2009 г. и отражает состояние лексики на начало XXI в.

Частотный словарь как особый тип лексикографического издания представляет собой одноязычный словарь, где приводятся лексемы данного языка с числовыми показателями, которые характеризуют употребительность лексем в каком-либо тексте – начиная от одиночного текста до собрания сочинений какого-либо автора (например, А.С. Пушкина, Ф.М. Достоевского, М.И. Цветаевой) или корпуса текстов одного или нескольких функциональных стилей. Одним из наиболее известных частотных

Беляков Михаил Васильевич – кандидат филологических наук, доцент кафедры русского языка Московского государственного института международных отношений (университета) МИД России.

Belyakov Mikhail Vasilievich – Doctor (Philology), Associate Professor, Department of the Russian Language, Moscow State Institute of International Relations (MGIMO University), Ministry of International Affairs.

E-mail: belmax007@hotmail.com

© Беляков М.В., 2021

словарей русского языка, полученным компьютерными средствами на большом материале (1 млн словоупотреблений), где были приведены частоты 40 тыс. слов, был словарь, созданный еще в 1977 г. под редакцией Л.Н. Засориной [3]. Поскольку современные частотные словари составляются по большей части на основании объемных корпусов, то для вычисления частотности был принят показатель *ipm* – количество словоупотреблений на миллион, т.е. стандартное употребление частоты леммы, вычисляемое относительно условного корпуса в миллион единиц независимо от объема реального корпуса. Как правило, очевидно наблюдаемым является отношение «ранг–частота», установленное еще в середине XX в. Дж. Ципфом, где ранг – это номер слова (леммы) в частотном списке.

Безусловно, объем выборки влияет на точность результатов. Результаты, полученные на небольшом по объему тексте, нельзя экстраполировать на весь язык, равно как и исследования на так называемых гигакорпусах, типа корпуса СМИ китайского языка объемом более миллиарда словоупотреблений, где будет ожидаемый перекоп в сторону тематической лексики.

Алгоритм исследования состоял в следующем: с использованием открытого компьютерного ресурса семантического анализа текста Istio (<https://istio.com/rus/text/analyz/>) последовательно проводился анализ текстов разных функциональных стилей с последующим сопоставлением имеющихся количественных характеристик частот в «Новом частотном словаре русской лексики».

2. Количественные характеристики функциональных стилей

Известно, что в русском языке традиционно выделяют пять функциональных стилей – художественный, научный, официально-деловой, публицистический и разговорный. Рассмотрим, какие количественные изменения относительно частоты встречаемости лексики произошли в некоторых из них.

Художественный стиль. Возьмем первые 20 слов по частоте встречаемости из словаря значимой лексики художественного стиля и посмотрим, как изменился этот список за период 2009–2020 гг. (табл. 1). Текстовый материал представляет собой выборку из нескольких книг, написанных в рамках художественного стиля и изданных в 2020 г. Однако необходимо отметить, что эти книги находятся в открытом доступе и написаны авторами-любителями, поэтому списки могут оказаться иными, если взять, например, образцы более «высокой» литературы. Принципиально важно для исследования, что выборка статистически релевантная и составляет 100 тысяч слов. Программный продукт семантического анализа Istio позволяет обрабатывать тексты такого объема (табл. 2).

Таблица 1

Словарь значимой лексики (художественный стиль)

№ п/п	Лемма	Часть речи	Частота в корпусе, ipm	Частота в подкорпусе, ipm
1	Ты	spro	3171,2	6153,6
2	Она	spro	5592,4	9048,2
3	Сказать	v	2396,6	4115,7
4	Он	spro	11791,1	17713,3
5	Спросить	v	573,9	1172,5
6	Рука	s	1200,6	2137,9
7	Глаз	s	857,6	1580,3
8	Голова	s	709,4	1326,7
9	Дверь	s	450,8	892,3
10	Подумать	v	315,6	626,4
11	Вдруг	adv	503,2	915,7
12	Плечо	s	236,4	477,5
13	Нога	s	459,2	834,6
14	Глядеть	v	199,3	410,3
15	Словно	conj	209,2	426,5
16	Улыбнуться	v	119,0	268,8
17	Кивнуть	v	88,4	214,1
18	Стоять	v	419,3	756,4
19	Палец	s	219,1	435,6
20	Губа	s	143,0	304,7

Таблица 2

Результат обработки выборки программой Istio

№ п/п	Слово	Кол-во	Релевантность	Процент в ядре	Процент в тексте
1	Рука	278	4,4	0,5	0,2
2	Даша	224	3,55	0,4	0,2
3	Голова	210	3,32	0,4	0,2
4	Глаз	185	2,93	0,3	0,1
5	Дом	140	2,21	0,2	0,1
6	Лицо	132	2,09	0,2	0,1
7	Говорить	131	2,08	0,2	0,1
8	Делать	122	1,93	0,2	0,1
9	Девушка	121	1,91	0,2	0,1
10	Мужчина	120	1,9	0,2	0,1
11	Нога	120	1,9	0,2	0,1
12	Взгляд	120	1,9	0,2	0,1
13	Палец	118	1,87	0,2	0,1
14	Стоять	118	1,87	0,2	0,1
15	Слово	116	1,87	0,2	0,1
16	Время	116	1,83	0,2	0,1
17	Жизнь	114	1,8	0,2	0,1
18	Дверь	114	1,8	0,2	0,1
19	Йорка	108	1,71	0,2	0,1
20	Человек	106	1,68	0,2	0,1

Стоит отметить, что стоп-слова (служебные слова, местоимения, ряд глаголов и пр.) не отображаются из-за специфики работы сайта. Рассмотрим совпадающие слова в обоих списках: *рука, глаз, голова, палец, дверь*. В списках совпадают 6 слов, причем, если принять во внимание проблемы с

подсчетом числа глаголов и местоимений, относящихся к стоп-словам, видно, что лексема *рука* как в словаре, так и по результатам исследования является наиболее употребляемым словом.

В обоих списках присутствуют слова, связанные с человеком, например, части тела, глаголы мышления и говорения и, что неожиданно, слово *дверь*. Скорее всего, это можно объяснить тем, что в современной художественной литературе центральной темой является человек, его поведенческие характеристики, его размышления. В 2020 г. частотны такие лексемы, как *мужчина* и *девушка*, хотя в Словаре они находятся на 61-м и 325-м месте соответственно. Также интересен тот факт, что 15-е, 16-е и 17-е места занимают слова: *слово*, *время*, *жизнь*, хотя они отсутствуют в словаре значимой лексики художественной литературы, и, что показательно, слово *человек* тоже не отмечено среди частотных в этом словаре.

Публицистический стиль. Как и в предыдущем случае, были отобраны первые 20 самых частотных слов выборки.

Для анализа была сформирована выборка текстов, взятых с нескольких новостных сайтов: РИА Новости, BBC Россия, Лента и новости портала Яндекс-новости за первое полугодие 2020 г. (табл. 3).

Таблица 3

**Словарь значимой лексики
(публицистический стиль)**

№ п/п	Лемма	Часть речи	Частота в корпусе, ipm	Частота в подкорпусе, ipm
1	Президент	s	311,0	634,6
2	Театр	s	305,3	611,1
3	Год	s	3727,5	5589,5
4	Спектакль	s	164,7	350,0
5	Правительство	s	277,7	531,2
6	Компания	s	392,7	699,0
7	Страна	s	725,7	1203,9
8	Фильм	s	196,8	380,1
9	Реформа	s	133,1	273,0
10	Выборы	s	117,7	243,4
11	Партия	s	250,4	448,7
12	Актер	s	114,4	231,6
13	Режиссер	s	103,5	211,1
14	Политический	adj	269,6	474,1
15	Советский	adj	382,9	643,9
16	Депутат	s	180,7	217,0
17	Известие	s	91,0	12,2
18	Лидер	s	92,2	186,4
19	Бюджет	s	126,0	241,1
20	Бизнес	s	136,2	256,7

Выборка также составляет 100 тысяч слов. Рассмотрим, как изменился этот список за прошедшее десятилетие (табл. 4). Новости по теме *коронавирус*

намеренно опускались, поскольку эта тема существенно изменила структуру новостей за последние три месяца, что мешает сравнению двух списков, так как эта тематика превратилась в абсолютно частотную.

Таблица 4

Результат обработки выборки программой Istio

№ п/п	Слово	Кол-во	Релевантность	Процент в ядре	Процент в тексте
1	Год	590	9,51	1	0,6
2	Человек	411	6,62	0,7	0,4
3	Время	248	3,99	0,4	0,2
4	Коронавирус	200	3,22	0,3	0,2
5	Страна	176	2,83	0,3	0,1
6	РОССИЯ	173	2,78	0,3	0,1
7	Женщина	162	2,61	0,2	0,1
8	Власть	147	2,36	0,2	0,1
9	Работа	145	2,33	0,2	0,1
10	Говорить	130	2,09	0,2	0,1
11	Мир	128	2,06	0,2	0,1
12	Фото	126	2,03	0,2	0,1
13	Слово	126	2,03	0,2	0,1
14	Тысяча	122	1,96	0,2	0,1
15	Мая	120	1,93	0,2	0,1
16	Работать	117	1,88	0,2	0,1
17	Компания	116	1,86	0,2	0,1
18	Российский	114	1,83	0,2	0,1
19	Жизнь	112	1,8	0,2	0,1
20	Дом	110	1,77	0,1	0,1
21	Пандемия	102	1,64	0,1	0,1

Результат удивил тем, что даже с учетом почти полного игнорирования новостных сводок о коронавирусе это слово, тем не менее, получило 4-е место, а пандемия – 21-е. При копировании материала возможны повторения слов, поэтому этот список был расширен (табл. 5).

Таблица 5

Результат обработки расширенной выборки

№ п/п	Слово	Кол-во	Релевантность	Процент в ядре	Процент в тексте
22	Война	102	1,64	0,1	0,1
23	Президент	97	1,56	0,1	0
24	Ребенок	95	1,53	0,1	0
25	Вопрос	94	1,51	0,1	0
26	Ситуация	94	1,51	0,1	0
27	Проблема	92	1,48	0,1	0

Рассмотрим слова, которые совпали в обоих списках.

1. *Год*. В списке, полученном в результате обработки выборки, это слово занимает первое место, в словаре – третье.

2. *Президент* – 23-е место в полученных результатах, первое место в словаре.

3. *Компания* – 17-е место в сформированном списке, шестое место в словаре.

Проблема сравнения выборки новостей состоит в том, что выборка такого рода очень чувствительна к хронологии ее формирования. В период первого этапа пандемии новостные сводки стали отличаться от тех, какими они были ещё несколько месяцев до неё. Все новости так или иначе концентрируются вокруг одной наиболее острой темы. По прошествии времени с изменением экстралингвистической ситуации диапазон тем заметно расширяется, открывая доступ новым переменным к списку частотных лексем.

Структура первой тысячи самых частых слов словаря также изменилась. Так, в сформированный по результатам компьютерной обработки выборки частотный список вошло слово *женщина*, которого нет в первой тысяче словаря. В нашем списке присутствует глагол *говорить* (10-е место), хотя в словаре среди первых 100 слов глаголы отсутствуют вообще. Также отметим, что слово *власть* стало более популярным, чем слово *правительство*. Скорее всего, за последние 10 лет произошел определенный семантический сдвиг в мировосприятии российского общества, так как эти слова приобрели отчасти отрицательную коннотацию.

Разговорный стиль. Современный видеохостинг Youtube предлагает возможность автоматического создания субтитров, что позволяет собирать и анализировать материал устной речи. Сформированная выборка состоит из более чем 40 различных видео на разную тематику. Это в основном подкасты и различные видеointервью. Стоит отметить, что правильное распознавание слов происходит только в 80 % случаев (табл. 6).

Таблица 6

Результат обработки выборки программой Istio

№ п/п	Слово	Кол-во	Релевантность	Процент в ядре	Процент в тексте
1	Тип	194	7,04	3	0,9
2	Человек	152	5,52	2,4	0,7
3	XXX	123	4,46	1,9	0,6
4	Говорить	88	3,19	1,3	0,4
5	Условный	59	2,14	0,9	0,2
6	Смотреть	37	1,34	0,5	0,1
7	Шутка	36	1,3	0,5	0,1
8	Жизнь	35	1,27	0,5	0,1
9	Момент	34	1,23	0,5	0,1
10	Время	30	1,08	0,4	0,1
11	Чувак	30	1,08	0,4	0,1
12	Вещь	29	1,05	0,4	0,1
13	XXX	28	1,01	0,4	0,1
14	Корона	28	1,01	0,4	0,1
15	История	27	0,98	0,4	0,1
16	XXX	26	0,94	0,4	0,1
17	Работать	25	0,9	0,3	0,1
18	Заниматься	25	0,9	0,3	0,1
19	Прямой	24	0,87	0,3	0,1
20	Год	23	0,83	0,3	0,1

Общечная лексика была скрыта значком (XXX), хотя ее присутствие стало неизбежной особенностью современного интернет-общения.

Для проверки полученных тенденций был проведен дополнительный эксперимент другой программой SEO-анализа – Advego (advego.ru). Для получения данных по частоте встречаемости выбранных лексем в данных текстах в программу загружаются тексты, а затем она производит все расчеты. Принципы работы данной системы аналогичны основным принципам статистической обработки системой Istio.

Для проведения эксперимента было отобрано 10 слов разных частей речи:

- вода (*сущ.*),
- правильный (*прил.*),
- знать (*гл.*),
- никак (*нареч.*),
- один (*числ.*),
- свободный (*прил.*),
- часть (*сущ.*),
- все (*местоим.*),
- сильный (*прил.*),
- несколько (*числ.*).

и 4 текста разных функциональных стилей:

1) *художественный стиль*: Щерба Н.В. «Часовое имя» (14 655 слов);

2) *публицистический стиль*: Фигерес К., Андерсон К. «Countdown: как нам повернуть вспять климатические изменения» (8076 слов);

3) *спонтанная устная речь*: Дудь Ю. «ВИЧ в России» (видеохостинг youtube.com, канал вДудь) (14 612 слов);

4) *научный стиль*: Гаджиев А.А., Хмелёв А.К. «Алгоритм Леска и система Babelfy для дизамбигуации»; Валуйцева И.И., Филатов И.Е. «Подходы к языковому и акустическому моделированию в распознавании речи» (9321 слово) (только символы кириллицы, без формул).

С помощью поиска по странице находились слова в списках частотности, сгенерированных приложением. В некоторых текстах отсутствовали определенные слова либо их показатель встречаемости был настолько мал, что система округляла его до нуля – Advego показывает результаты в процентах с двумя символами после запятой.

Затем были определены значения частотности лексем в Новом частотном словаре (НЧС) [2]. В списках частотного словаря значения представлены в единицах ipm (количество употреблений на миллион слов), поэтому значения были переведены в проценты. Полученные данные были внесены в табл. 7. Такое представление результатов эксперимента удобно для выявления изменений частотности.

Таблица 7

Результат обработки выборки программой Advego

Слово	Частота, %							
	Художественный	Художественный (НЧС)	Публицистический	Публицистический (НЧС)	Научный	Научный (НЧС)	Устная речь	Устная речь (НЧС)
Вода	0.01	0.06	0.00	0.04	0.00	0.05	0.00	0.03
Правильный	0.01	0.01	0.04	0.01	0.09	0.01	0.01	0.01
Знать	0.29	0.25	0.09	0.13	0.02	0.10	0.32	0.63
Никак	0.01	0.01	0.00	0.01	0.02	0.01	0.01	0.02
Один	0.29	0.22	0.35	0.25	0.48	0.20	0.37	0.23
Свободный	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.01
Часть	0.02	0.03	0.05	0.07	0.20	0.11	0.02	0.02
Все	0.66	0.40	0.80	0.29	0.30	0.17	0.84	0.57
Сильный	0.10	0.02	0.04	0.02	0.00	0.20	0.04	0.01
Несколько	0.07	0.05	0.02	0.08	0.27	0.50	0.12	0.02

Выводы

Значения частотности слов в художественной литературе за рассматриваемый период значительно не изменились – временной диапазон слишком мал для вывода о том, что в частоте встречаемости данных лексем произошли перемены.

В публицистических текстах изменения куда более заметны – почти все лексемы, кроме *знать*, *часть* и *несколько*, употребляются в современном тексте реже. Однако можно заметить, что высокочастотные слова словаря [2] (*один*, *все*) остаются таковыми и в тексте нашей выборки.

Сравнение частоты лексем в научном тексте с показателями из частотного словаря нехудожественной литературы показывает существенную разницу в значениях. Однако так же, как в публицистической литературе, частотность определенных слов остается высокой в обоих текстах.

Значения частотности слов живой устной речи изменились не сильно, но при этом динамика изменения значений частотности для высокочастотных слов также схожа с публицистическим стилем. Редко встречаемые слова сохраняют свои показатели.

Заключение

Анализ полученных результатов выявляет некоторые неожиданные факты. Изменилась структура частотных списков – в Словаре [2] представлены в первую очередь имена существительные, в сформированных компьютерными программами Istio и Advego списках частеречный диапазон шире, в них появились глаголы. Так, в выборке публицистического материала присутствуют слова *работать* и *говорить*; оказалось, что слово *условный* употребляется в живой речи с очень высоким частотным показателем, а слова *жизнь*, *момент*, *время* часто употребляются в современных текстах художественного стиля. К сожалению, не представляется возможным проанализировать частотный список стоп-слов (сайт не справляется с таким объемом). В устной речи это бы могло показать интересные результаты.

Тем не менее выводы по сравнительному лингвостатистическому анализу экспериментальных данных и данных «Нового частотного словаря русской лексики» [2] показывают, что, несмотря на уже устоявшееся мнение, что в русском языке в последние десятилетия произошел лексический взрыв из-за активного заимствования слов вместе с денотатами из английского языка (чаще его американского варианта), активного использования ранее малопродуктивных префиксальных моделей словообразования (префиксоиды *анти-*, *пост-*, *контр-* и др.), аббревиатур [1], семантических сдвигов в значении слов, возвращения в узус слов, ранее находившихся на лексической периферии (*кадет*, *губернатор*, *светский*, *безработица*) и др., эти слова не попадают в список наиболее частотных слов разных функциональных стилей. Единственным исключением является обценная лексика, проникшая как в художественный стиль (в нашей выборке это явление не отражено), так и в публицистический стиль (в широком понимании), а именно в интернет-коммуникацию, превратившись в один из ее маркеров.

Таким образом, с одной стороны, можно говорить об ускоренном хронологическом шаге (см. выше) лексических изменений в русском языке, ощущаемых на интуитивном уровне, а с другой, о достаточной стабильности высокочастотной лексики всех функциональных стилей, что показало данное лингвостатистическое исследование.

ЛИТЕРАТУРА

1. Максименко, О.И. Новые тенденции в аббревиации (на материале русского, английского и немецкого языков) / О.И. Максименко // Вестник РУДН. – 2017. – Т. 8, № 1. – С. 174–182. – (Сер. Теория языка. Семиотика. Семантика).
2. Новый частотный словарь русской лексики (на материалах Национального корпуса русского языка) / сост. О.Н. Ляшевская, С.А. Шаров. – Москва : Азбуковник, 2009. – URL: <http://dict.ruslang.ru/freq.php> (дата обращения: 01.12.2020).
3. Частотный словарь русского языка / под ред. Л.Н. Заориной. – Москва : Русский язык, 1977. – 936 с.