Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2016»

Москва, 1–4 июня 2016

НАИБОЛЕЕ УПОТРЕБИТЕЛЬНЫЕ СЛОВА ПОВСЕДНЕВНОЙ РУССКОЙ РЕЧИ (В ГЕНДЕРНОМ АСПЕКТЕ И В ЗАВИСИМОСТИ ОТ УСЛОВИЙ КОММУНИКАЦИИ)[1]

Шерстинова Т. Ю. (t.sherstinova@spbu.ru)

СПбГУ, Санкт-Петербург, Россия

В центре внимания данной работы находятся наиболее употребительные слова русской повседневной речи, представляющие собой верхнюю зону частотных словарей, полученных на материале звукозаписей речевого корпуса «Один речевой день» (ОРД). Речевой материал, представленный в корпусе, проаннотирован с точки зрения условий коммуникации (тип коммуникации/языковой стиль, социальная роль говорящего, локус и др.), а также снабжен информацией о социальных характеристиках информанта и его основных коммуникантов. Такая информация позволяет фильтровать речевой материал и исследовать изменение речевых характеристик в зависимости от социальных характеристик говорящего и условий коммуникации. Исследование выполнено на материале 152 эпизодов повседневной речевой коммуникации, которые содержат в общей сложности 232 370 словоупотреблений. Выборка содержит речь 209 говорящих (95 мужчин, 94 женщины, 20 детей). Построены общий частотный словарь, мужской и женские словари устной речи, а также гендерные словари для четырех стилей устной речи: 1) бытовые неформальные разговоры, 2) профессиональные (деловые/официальные) разговоры, 3) учебно-образовательная речевая коммуникация, 4) коммуникация по типу «клиент-сервис».

Ключевые слова: русская разговорная речь, повседневная речевая коммуникация, частотные словари, социолингвистика, гендерная вариативность речи, стили устной речи, речевой корпус, устный дискурс

---

[1] Исследование выполнено при поддержке гранта РНФ № 14-18-02070 «Русский язык повседневного общения: особенности функционирования в разных социальных группах».

THE MOST FREQUENT WORDS IN EVERYDAY SPOKEN RUSSIAN (IN THE GENDER DIMENTION AND DEPENDING ON COMMUNICATION SETTINGS)

Sherstinova T. Yu. (t.sherstinova@spbu.ru)

Saint-Petersburg State University, Saint-Petersburg, Russia

The paper presents the most frequent words of everyday spoken Russian, that form the upper zones of several word frequency lists compiled on the material of Russian speech corpus “One Speaker’s Day” (the ORD corpus), containing real-life recordings of everyday communication. All speech data in the corpus is annotated in terms of communication settings, including 1) type of communication (language spoken style), 2) social role of speaker, 3) locus, etc. Such information allows speech to be filtered upon user request and therefore makes it possible to study speech variation depending on particular communication settings. The given study was made on the transcripts of 152 real-life macroepisodes and contains 232,370 words. The sample presents speech of 209 persons (95 men, 94 women, 20 children). The following word frequency lists have been compiled: a) general frequency list, b) male frequency list, c) female frequency list, and d) four frequency lists for different styles of spoken speech: informal conversations, professional/business conversations, educational communication, and “customer-service” communication. Men’s and women’s frequency lists have been compiled on the subsamples of 83,371 and 115,110 words correspondently. The analysis of word lists has shown that Russian women pay more attention to maintaining the conversation, use fewer hesitations, and are more inclined to use in their speech intensifying words, emotional words, hedges and interjections. Men generally use fewer personal pronouns, while numbers and the expletives are among the most frequent words used by men in everyday conversations. In general, these observations are similar to those described earlier for gender variation by other linguists.

Key words: spoken Russian, everyday verbal communication, word frequency lists, sociolinguistics, gender variability of speech, styles of spoken language, speech corpus, oral discourse

1. Введение

Построение частотных словарей — традиционный метод современных лексикографических исследований на базе лингвистических корпусов (см., например, [Мартыненко 1988; Leech et. al 2001; Popescu 2009; Ляшевская, Шаров, 2009; Шайкевич 2015] и др.). Частоту языковых единиц можно рассматривать как индикатор маркированности [Baker, 2010, p. 125], который позволяет оценить их функциональную активность. Данные, представленные в «сухой» табличной форме частотного словаря, лаконично и по существу выявляют

особенности текстового материала и позволяют проводить его статистический анализ. Особый интерес исследователей привлекает «верхняя зона» частотных словарей, представляющая костяк речевой системы, во многом совпадающая не только для индивидуальных говорящих, но и для целых подсистем (языковых стилей, регистров, ситуаций общения). Для выявления таких зон предлагаются специальные индексы [Мартыненко 1988, Popescu 2009].

За последние десятилетия подготовлено большое количество частотных словарей как для отдельных языковых жанров (научной и художественной литературы, специальных текстов), так и для отдельных выдающихся писателей. Однако до сих пор лингвисты не располагают статистически представительными данными о частоте слов в наиболее важном для человека языковом жанре — живой спонтанной речи, которая составляет основу повседневной коммуникации. Предлагаемое исследование позволяет отчасти восполнить этот пробел, представляя частотные списки словоформ в зависимости от разных стилей повседневной устной речи в речи мужчин и женщин, в разных ситуациях общения.

Другой целью данной работы является демонстрация возможностей речевого корпуса ОРД, предоставляющего возможность выборки и анализа данных не только по социодемографическим характеристикам говорящего и стилю устной речи, являющихся традиционными для многих речевых корпусов, но также и в зависимости от разных условий коммуникации [Sherstinova 2015].

2. Речевой корпус ОРД

Разработка речевого корпуса повседневной русской речи ОРД («Один речевой день»), для получения звукозаписей которого информанты-добровольцы согласились прожить целый день с «диктофоном на шее», записывающим всю их речевую коммуникацию, позволяет вывести исследования повседневной устной речи на качественно новый уровень [Asinovsky et al. 2009]. Подобная методика записи речи традиционно используется в японских лингвистических исследованиях (см., например, [Campbell 2004]), а также применялась при подготовке материалов для устного подкорпуса BNC [Burnard 2007]. Ее преимущество состоит в получении для анализа речевого материала, наиболее приближенного к естественной повседневной речи.

В настоящее время корпус содержит более 1200 часов звукозаписей речи, полученной от 127 информантов, мужчин и женщин, в возрасте от 17 до 77 лет, и нескольких сотен их коммуникантов. Сбор материала и аннотирование корпуса продолжается [Bodganova-Beglarian et al. 2015].

Все звукозаписи «речевого дня» подлежат сегментации на макроэпизоды повседневного общения, которые аннотируются с точки зрения условий коммуникации (локуса, социальных ролей участников, типа коммуникации и некоторых ее особенностей) [Sherstinova 2015]. Нормализация соответствующих кодов позволяет использовать эти параметры в качестве фильтров для выборки необходимого речевого материала. Благодаря этим фильтрам были получены представленные в работе списки частотных слов.

3. Материал и методика

Данное разведочное исследование выполнено на расшифровках 152 макроэпизодов, записанных 40 информантами и их коммуникантами. Записи выполнены в Санкт-Петербурге в 2007 и 2010 гг. Выборка содержит речь 209 человек:

Информанты		**Основные коммуниканты**				**Итого (человек)**
Муж.	**Жен.**	**Муж.**	**Жен.**	**Дети**	**Всего**	
20	20	75	74	20	169	209

Общая длительность звучания проанализированного речевого материала составляет 40,5 часов, в среднем по 3,8 эпизода и 1 часу звукозаписи от каждого информанта. Объем подкорпуса в словоупотреблениях составляет 232 370 единиц. Средняя продолжительность эпизода — 15,99 мин. (SD = 9,80 мин.), в словах — 1550 (SD = 1120).

По типам коммуникации, коррелирующим с соответствующими стилями устной речи, эпизоды исследуемой выборки имеют следующее распределение:

Тип коммуникации / Стиль устной речи	**Процент эпизодов**	**Кол-во эпизодов**	**Кол-во словоупотр.**
Бытовые неформальные разговоры	62,5%	95	154 051
Профессиональные (деловые) разговоры	19,7%	30	40 012
Коммуникация по типу «клиент-сервис»	9,9%	15	21 126
Учебная коммуникация	7,2%	11	19 629
Публичные выступления	0,7%	1	800

Учебная коммуникация понимается как общение между «обучающим» и «обучаемым» (лекции, практические занятия, индивидуальные занятия, инструктаж, обучающее занятие с ребенком и т. п.). Коммуникация по типу «клиент-сервис», как правило, представляет собой относительно формальный разговор между человеком, профессионально оказывающим «услуги» в широком смысле слова (кроме обучающих) и его клиентом (в государственных службах, сервис-центрах, магазинах, поликлиниках, библиотеках и т. п.). Профессиональная беседа — деловой разговор на профессиональные темы, не являющийся ни учебной, ни «клиент-сервисной» коммуникацией, осуществляемый преимущественно между коллегами. Наконец, «бытовой разговор» имеет неформальный характер, в большинстве случаев не связан тематически с профессиональной деятельностью информанта и может иметь место среди разных коммуникантов, в разных условиях общения. Как любое прагматическое аннотирование, атрибуция макроэпизодов по типам коммуникации в большой степени зависит от контекста и содержания беседы.

Полученное распределение в целом согласуется с распределением эпизодов по типу коммуникации, посчитанных для корпуса ОРД на материале 1854 макроэпизодов, описывающих 483 часов звучащей речи [Sherstinova 2015].

Следует отметить следующие особенности построения частотных списков:

1) В связи с тем, что процесс лемматизации и разведения омонимии корпуса ОРД еще не завершен, здесь анализируются словоформы, без учета омонимичных форм. Впрочем, словоформы довольно часто используются при анализе частот [Rayson et al. 1997; Popescu 2009]. В нашем случае «смягчающим» обстоятельством, оправдывающим такой подход, является тот факт, что рассматриваемая нами верхняя зона частотного словаря состоит преимущественно из неизменяемых частиц. Однако очевидно, что при лемматизации следует ожидать роста частот личных местоимений — другой высокоактивной категории единиц.
2) Все слова в транскриптах ОРД записаны в стандартной орфографии, независимо от реального произнесения слова («*сейчас*», а не «*щас*»; «*что*», а не «*чё*» или «*шо*»), поскольку особенности фонетической реализации слов в корпусе ОРД отмечаются на специальных фонетических уровнях.
3) В представленных частотных списках было решено оставить и некоторые «несловарные» элементы устной речи (в частности, «*угу*» и «*ага*», а также заполненные (э) и незаполненные (…) паузы хезитации).
4) При интерпретации данных следует иметь в виду, что двухсловные языковые единицы, такие как «*потому что*», «*так как*» и т. п., как и в большинстве других частотных словарей, представлены здесь отдельными «частями».
5) В качестве «наиболее употребительных» слов рассматривается первая сотня словоформ верхней зоны соответствующего частотного словаря.

4. Общий частотный словарь

Прежде всего рассмотрим распределение наиболее активной лексики в целом по выборке при отсутствии каких-либо фильтров.

В табл. 1 каждое слово сопровождается численной информацией: ранг слова, его абсолютная частота по выборке, доля (в процентах), а также кумулятивный процент. Последний показатель довольно интересен, так как позволяет оценить совокупную долю частот с рангом не ниже данной. Так, из табл. 1 видно, что 10 наиболее употребительных слов устной речи (*я*, *вот*, *ну*, *не*, *да*, *а*, *и*, *что*, *в*, *это*) покрывают примерно 1/5 (20,69%) всей речевой коммуникации, а 88 самых частотных слов покрывают уже половину (50,00%) всего речевого материала.

Наиболее употребительными словами устной русской речи на нашем материале оказались личные местоимения (*я*, *ты*, *он*, *она*, *они*, *мы*, *вы* в форме им. п., *мне* (дат./предл.), *меня* (род./вин. п.)) и неизменяемые единицы — частицы, союзы, предлоги и слова, которые в последнее время все чаще стали называть «дискурсивными словами» или «маркерами» (*вот*, *ну*, *не*, *да*, *нет*, *так*, *просто, вообще*, *значит*, *сейчас*, и др.). Многие из этих единиц выполняют в речи прагматические функции и/или используются для регулирования дискурса. Среди частотных мы видим и заполненную хезитацию (*э*). Другая часть высокочастотных слов репрезентирует кластеры омонимичных форм, аттрибутировать которые невозможно без обращения к контексту (*а*, *есть*, *что* и др.).

Необходимо отметить, что полученные данные в определенной степени коррелируют с другими частотными словарями устной речи, подготовленными на материале русского [Ляшевская, Шаров 2009] и английского языков [Leech et al. 2001], но и имеют свою специфику.

Таблица 1. Наиболее употребительные слова устной русской речи (общий список)

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
1	я	5950	2,56	2,56	51	очень	553	0,24	42,97
2	вот	5411	2,33	4,89	52	все	550	0,24	43,21
3	ну	5407	2,33	7,22	53,5	потом	545	0,23	43,45
4	не	5400	2,32	9,54	53,5	тебе	545	0,23	43,68
5	да	5162	2,22	11,76	55	говорит	542	0,23	43,91
6	а	4523	1,95	13,71	56	может	536	0,23	44,14
7	и	4190	1,80	15,51	57	хорошо	532	0,23	44,37
8	что	4139	1,78	17,29	58	когда	527	0,23	44,60
9	в	4003	1,72	19,01	59	или	525	0,23	44,83
10	это	3905	1,68	20,69	60	его	496	0,21	45,04
11	там	3422	1,47	22,17	61	такой	489	0,21	45,25
12	у	3007	1,29	23,46	62	тебя	485	0,21	45,46
13	так	2842	1,22	24,68	63	за	480	0,21	45,66
14	на	2602	1,12	25,80	64	давай	468	0,20	45,87
15	как	2110	0,91	26,71	65	говорю	465	0,20	46,07
16	ты	1877	0,81	27,52	66	только	464	0,20	46,27
17	всё	1874	0,81	28,33	67	ой	463	0,20	46,47
18	то	1781	0,77	29,09	68	можно	450	0,19	46,66
19	с	1771	0,76	29,85	69,5	потому	449	0,19	46,85
20	нет	1727	0,74	30,60	69,5	знаешь	449	0,19	47,05
21	(э)	1708	0,74	31,33	71	где	432	0,19	47,23
22	он	1670	0,72	32,05	72	бл**ь	430	0,19	47,42
23	угу	1451	0,62	32,68	73	ага	426	0,18	47,60
24	мне	1293	0,56	33,23	74	ничего	425	0,18	47,78
25	она	1275	0,55	33,78	75,5	конечно	405	0,17	47,96
26	есть	1256	0,54	34,32	75,5	что-то	405	0,17	48,13
27	меня	1124	0,48	34,81	77	этот	401	0,17	48,30
28	сейчас	1117	0,48	35,29	78	чего	390	0,17	48,47
29	они	1060	0,46	35,74	79	вас	377	0,16	48,63
30	мы	1050	0,45	36,19	80	такая	369	0,16	48,79
31	бы	1002	0,43	36,63	81	раз	366	0,16	48,95
32	но	989	0,43	37,05	82	такое	359	0,15	49,10
33	уже	984	0,42	37,47	83	до	357	0,15	49,26
34	надо	953	0,41	37,88	84	её	356	0,15	49,41
35	ещё	938	0,40	38,29	85	чтобы	343	0,15	49,56
36	же	914	0,39	38,68	86	вам	342	0,15	49,71
37	по	889	0,38	39,06	87	эти	341	0,15	49,85

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
38	просто	875	0,38	39,44	88	даже	340	0,15	50,00
39	вообще	772	0,33	39,77	89	их	334	0,14	50,14
40	если	721	0,31	40,08	90	блин	328	0,14	50,28
41	вы	712	0,31	40,39	91	был	321	0,14	50,42
42	нас	691	0,30	40,69	92	два	312	0,13	50,56
43	тоже	665	0,29	40,97	93	для	309	0,13	50,69
44	знаю	664	0,29	41,26	94	один	305	0,13	50,82
45	было	592	0,25	41,51	95	кто	303	0,13	50,95
46	значит	584	0,25	41,76	96	быть	300	0,13	51,08
47	здесь	577	0,25	42,01	97,5	из	299	0,13	51,21
48	будет	570	0,25	42,26	97,5	ладно	299	0,13	51,34
49	к	556	0,24	42,50	99	ли	297	0,13	51,47
50	тут	555	0,24	42,74	100	короче	296	0,13	51,59

5. Наиболее частотные слова мужской и женской речи

Исследование гендерной вариативности речи представляет собой популярное направление в российской и мировой лингвистике [Lakoff, 1975; Tannen, 1991; Потапова, Потапов 2006; и мн. др.]). Ее изучение на материале ОРД возможно благодаря сбалансированности информантов корпуса по гендерному параметру.

Так, мужская и женская речь представлены в анализируемой выборке практически одинаковым количеством говорящих, однако объем «женского подкорпуса» в словах превысил «мужской» примерно на треть, что дает основания предполагать, что мужчины в среднем говорят меньше. С другой стороны, индекс лексического богатства для речи мужчин оказался несколько выше, чем у женщин:

	Количество говорящих	Всего словоупотреблений[2]	Всего разных словоформ	Всего однократных форм	Индекс лексического богатства
Мужской подкорпус	95	83 371	14 539	9 408	0,174
Женский подкорпус	94	115 110	17 470	10 927	0,152

В табл. 2 и 3 приводятся наиболее употребительные слова женской и мужской устной русской речи соответственно.

Первое, что бросается в глаза, это отличие слов «первого» ранга. Личное местоимение *я*, являющееся абсолютным лидером в женской речи, уступает

[2] В гендерные подкорпуса по сравнению с общей выборкой не была включена речь детей и подростков до 18 лет, а также ряд фрагментов, характеризующиеся одновременной речью двух или более говорящих.

пальму первенства частице *ну*[3] в мужской речи. Видно, что женщины уделяют больше внимания формальному поддержанию разговора (*угу*, *хорошо*), меньше хезитируют и чаще используют усилительные слова (*очень*) и междометия (*ой*).

Мужчины реже употребляют в речи личные местоимения, у них чаще наблюдаются хезитации. Наиболее отличительной чертой мужской речи по сравнению с женской является отмечаемое всеми без исключениями исследователями [Lakoff 1975, Stenström 1991, Rayson et al. 1997, и др.] появление в верхней зоне частотного словаря бранной лексики, непечатных слов и их субститутов.

В связи с тем, что мужской и женский словари не совпадают по составу лексических единиц, в том числе и в «верхней зоне», сравнение частоты употребления проводится относительно или первого, или второго списка. Так, например, выглядят пятерки «преимущественно женских» и «мужских» слов:

«Наиболее женские» слова				**«Наиболее мужские» слова**			
По разности рангов		**По разности долей**		**По разности рангов**		**По разности долей**	
нам	68,5	угу	0,377%	х**	–	б**дь	0,504%
как-то	61,5	так	0,370%	б**дь	3365	в	0,464%
слушай	60,5	что	0,326%	блин	283	там	0,452%
вам	49	да	0,322%	короче	98,5	на	0,313%
ой	40	я	0,303%	типа	72,5	блин	0,311%

В целом наблюдаемые на материале ОРД различия в речи мужчин и женщин совпадают с выводами, сделанными ранее, в том числе и на материале других языков [Coupland 2007; Romaine 2008; Потапова, Потапов 2006; и др.].

Таблица 2. Наиболее употребительные слова женской устной русской речи

Ранг	**Слово**	**Абс. ч.**	**%**	**Кумул. %**	**Ранг**	**Слово**	**Абс. ч.**	**%**	**Кумул. %**
1	я	3110	2,70	2,70	51,5	знаешь	274	0,24	43,59
2	не	2722	2,36	5,07	51,5	все	274	0,24	43,83
3	вот	2691	2,34	7,40	53	говорит	271	0,24	44,06
4	да	2593	2,25	9,66	54	потом	268	0,23	44,29
5	ну	2577	2,24	11,90	55	тебе	262	0,23	44,52
6	что	2218	1,93	13,82	56	или	260	0,23	44,75
7	а	2175	1,89	15,71	57	может	259	0,23	44,97
8	и	2116	1,84	17,55	58	давай	255	0,22	45,19
9	это	1925	1,67	19,22	59	говорю	254	0,22	45,41
10	в	1770	1,54	20,76	60	здесь	246	0,21	45,63
11	так	1619	1,41	22,17	61	тут	245	0,21	45,84

[3] Проведенное ранее исследование позволяет предположить, что использование частицы «*ну*» более характерно для неформального общения, в отличие от частицы «*вот*», которая чаще используется в официальной (деловой) речи (Sherstinova 2015), однако эта гипотеза требует проверки.

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
12	у	1569	1,36	23,53	62	когда	241	0,21	46,05
13	там	1490	1,29	24,82	63	только	238	0,21	46,26
14	на	1176	1,02	25,85	64,5	его	236	0,21	46,46
15	как	1140	0,99	26,84	64,5	ничего	236	0,21	46,67
16	ты	1004	0,87	27,71	66	что-то	235	0,20	46,87
17	всё	932	0,81	28,52	67	такой	231	0,20	47,07
18	с	913	0,79	29,31	68	тебя	230	0,20	47,27
19	угу	836	0,73	30,04	69	где	223	0,19	47,47
20	то	826	0,72	30,75	70,5	значит	220	0,19	47,66
21	нет	798	0,69	31,45	70,5	конечно	220	0,19	47,85
22	она	783	0,68	32,13	72	вас	218	0,19	48,04
23	мне	749	0,65	32,78	73	потому	217	0,19	48,23
24	он	739	0,64	33,42	74	можно	214	0,19	48,41
25	(э)	677	0,59	34,01	75,5	за	208	0,18	48,59
26	мы	651	0,57	34,57	75,5	ага	208	0,18	48,77
27	сейчас	571	0,50	35,07	77	такая	198	0,17	48,95
28	меня	568	0,49	35,56	78	вам	197	0,17	49,12
29	есть	564	0,49	36,05	79	её	195	0,17	49,29
30	бы	561	0,49	36,54	80	такое	194	0,17	49,45
31	но	523	0,45	37,00	81	даже	171	0,15	49,60
32	они	518	0,45	37,45	82	ладно	167	0,15	49,75
33	надо	486	0,42	37,87	83,5	чтобы	166	0,14	49,89
34	ещё	477	0,41	38,28	83,5	чего	166	0,14	50,04
35	уже	469	0,41	38,69	85,5	этот	164	0,14	50,18
36	же	452	0,39	39,08	85,5	эти	164	0,14	50,32
37	по	430	0,37	39,46	87	их	162	0,14	50,46
38,5	вообще	423	0,37	39,82	88,5	нам	161	0,14	50,60
38,5	вы	423	0,37	40,19	88,5	сегодня	161	0,14	50,74
40	просто	407	0,35	40,54	90,5	раз	159	0,14	50,88
41	нас	395	0,34	40,89	90,5	до	159	0,14	51,02
42	тоже	364	0,32	41,20	92	как-то	158	0,14	51,16
43	знаю	352	0,31	41,51	93	для	157	0,14	51,29
44	очень	342	0,30	41,81	94	ли	152	0,13	51,42
45	если	326	0,28	42,09	95	о	148	0,13	51,55
46	к	298	0,26	42,35	96	них	147	0,13	51,68
47	будет	294	0,26	42,60	97	слушай	145	0,13	51,81
48	хорошо	292	0,25	42,86	98	пока	144	0,13	51,93
49	было	288	0,25	43,11	99,5	быть	143	0,12	52,06
50	ой	278	0,24	43,35	99,5	думаю	143	0,12	52,18

Таблица 3. Наиболее употребительные слова мужской устной русской речи

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
1	ну	2062	2,47	2,47	50,5	было	208	0,25	42,29
2	я	2000	2,40	4,87	52	потом	207	0,25	42,54
3	не	1898	2,28	7,15	53	все	205	0,25	42,78
4	вот	1887	2,26	9,41	54,5	короче	201	0,24	43,02
5	в	1669	2,00	11,41	54,5	тоже	201	0,24	43,26
6	да	1610	1,93	13,35	56	такой	194	0,23	43,50
7	а	1548	1,86	15,20	57	тебя	192	0,23	43,73
8	и	1530	1,84	17,04	58	будет	191	0,23	43,96
9	там	1456	1,75	18,78	59	нас	189	0,23	44,18
10	это	1338	1,60	20,39	60	или	183	0,22	44,40
11	что	1335	1,60	21,99	61	чего	182	0,22	44,62
12	на	1113	1,33	23,32	62	его	177	0,21	44,83
13	у	985	1,18	24,51	63	может	173	0,21	45,04
14	так	864	1,04	25,54	64	только	172	0,21	45,25
15	(э)	761	0,91	26,46	65	этот	169	0,20	45,45
16	он	696	0,83	27,29	66	потому	166	0,20	45,65
17	то	680	0,82	28,11	67	к	165	0,20	45,85
18	всё	665	0,80	28,90	68	вы	157	0,19	46,04
19	как	656	0,79	29,69	69	хорошо	156	0,19	46,22
20	ты	651	0,78	30,47	70	давай	153	0,18	46,41
21	с	609	0,73	31,20	71	раз	147	0,18	46,58
22	нет	537	0,64	31,85	72,5	один	146	0,18	46,76
23	есть	478	0,57	32,42	72,5	говорю	146	0,18	46,93
24	бл**ь	422	0,51	32,93	74	где	144	0,17	47,11
25	сейчас	410	0,49	33,42	75	можно	143	0,17	47,28
26	меня	393	0,47	33,89	76	кто	139	0,17	47,44
27	они	387	0,46	34,35	77,5	был	137	0,16	47,61
28	мне	356	0,43	34,78	77,5	знаешь	137	0,16	47,77
29	уже	344	0,41	35,19	79,5	ничего	134	0,16	47,93
30	просто	331	0,40	35,59	79,5	их	134	0,16	48,09
31	ещё	330	0,40	35,98	81	очень	133	0,16	48,25
32	но	325	0,39	36,37	82	два	132	0,16	48,41
33	надо	320	0,38	36,76	83	до	131	0,16	48,57
34	же	319	0,38	37,14	84,5	чтобы	127	0,15	48,72
35	она	318	0,38	37,52	84,5	такая	127	0,15	48,87
36	по	315	0,38	37,90	86	типа	126	0,15	49,02
37	значит	294	0,35	38,25	87	такое	125	0,15	49,17
38	угу	291	0,35	38,60	88,5	ага	124	0,15	49,32
39	блин	283	0,34	38,94	88,5	из	124	0,15	49,47
40	если	281	0,34	39,28	90	ой	122	0,15	49,62
41	бы	274	0,33	39,61	91	даже	121	0,15	49,76
42	вообще	269	0,32	39,93	92,5	что-то	119	0,14	49,91

Ранг	Слово	Абс. ч.	%	Кумул. %	Ранг	Слово	Абс. ч.	%	Кумул. %
43	знаю	248	0,30	40,23	92,5	конечно	119	0,14	50,05
44	мы	228	0,27	40,50	94,5	её	118	0,14	50,19
45,5	тут	218	0,26	40,76	94,5	х**	118	0,14	50,33
45,5	говорит	218	0,26	41,02	96	три	117	0,14	50,47
47	здесь	215	0,26	41,28	97	эти	115	0,14	50,61
48	за	213	0,26	41,54	98	для	110	0,13	50,74
49	когда	210	0,25	41,79	99	понятно	108	0,13	50,87
50,5	тебе	208	0,25	42,04	100	тогда	107	0,13	51,00

В качестве примера того, как может перераспределяться функциональная активность слов в зависимости от коммуникативной ситуации, приведем 2 таблицы, в которых показаны по 50 наиболее частотных слов тех же мужского и женского подкорпусов для разных языковых стилей[4]: 1) неформальной бытовой речи, 2) профессионального разговора, 3) «учебной» коммуникации и 4) широкого пласта коммуникативных ситуаций по типу «клиент-сервис».

Поскольку при дроблении выборки на подкатегории статистическая представительность каждой из них неизбежно уменьшается, представленные данные носят скорее иллюстративный характер (см. табл. 4 и 5).

	Объем подкорпусов разных стилей (словоупотребления)			
	Бытовой	Профессиональный	«Клиент-сервис»	Учебный
Женский подкорпус	72 759	25 239	6 838	10 094
Мужской подкорпус	62 030	10 123	5 187	5 254

Таблица 4. Наиболее употребительные слова женской устной речи в зависимости от коммуникативной ситуации (стиля речи)

Бытовая			Профессиональная			Образовательная			«Клиент-сервис»		
Ранг	Слово	%	Ранг	Слово	%	Ранг	Слово	%	Ранг	Слово	%
1	я	2,87	1	я	2,92	1	да	3,35	1	вот	3,01
2	не	2,61	2	вот	2,59	2	вот	3,05	2	я	2,28
3	ну	2,46	3	что	2,27	3	ну	1,85	3	не	2,19
4	да	2,23	4	не	2,21	4	так	1,80	4	и	1,97
5	а	2,17	5	и	2,11	5	это	1,63	5	да	1,97
6	вот	2,09	6	да	1,96	6	что	1,60	6	это	1,95
7	что	1,88	7	ну	1,87	7	у	1,43	7	а	1,87
8	и	1,81	8	в	1,75	8	и	1,31	8	ну	1,81
9	это	1,66	9	это	1,65	9	я	1,24	9	что	1,59

[4] Подробнее о стилях см. п. 3.

Бытовая			Профессиональная			Образовательная			«Клиент-сервис»		
Ранг	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**
10	в	1,57	10	а	1,44	10	не	1,14	10	угу	1,52
11	там	1,40	11	там	1,34	11	мы	1,06	11	в	1,43
12	так	1,40	12	так	1,30	12	вы	1,01	12	у	1,36
13	у	1,40	13	у	1,23	13	угу	1,01	13	так	1,29
14	на	1,16	14	(э)	1,13	14	а	1,00	14	всё	1,10
15	ты	1,08	15	то	1,05	15	всё	0,92	15	как	1,10
16	как	1,02	16	всё	1,01	16	там	0,89	16	то	0,98
17	с	0,83	17	как	0,92	17	по	0,89	17	есть	0,94
18	он	0,74	18	она	0,90	18	как	0,86	18	на	0,88
19	нет	0,73	19	на	0,82	19	в	0,84	19	вам	0,85
20	всё	0,70	20	с	0,78	20	(э)	0,70	20	с	0,78
21	мне	0,68	21	есть	0,74	21	на	0,65	21	она	0,76
22	она	0,67	22	мне	0,73	22	с	0,57	22	вы	0,75
23	угу	0,63	23	угу	0,66	23	нас	0,56	23	нет	0,73
24	то	0,63	24	нет	0,65	24	сейчас	0,51	24	сейчас	0,72
25	мы	0,57	25	ты	0,57	25	вас	0,50	25	бы	0,67
26	они	0,53	26	меня	0,56	26	давай	0,49	26	мне	0,63
27	бы	0,53	27	сейчас	0,53	27	нет	0,47	27	(э)	0,60
28	меня	0,52	28	вы	0,53	28	здесь	0,47	28	мы	0,58
29	но	0,48	29	но	0,47	29	он	0,44	29	он	0,57
30	сейчас	0,46	30	он	0,46	30	можно	0,42	30	вас	0,57
31	вообще	0,45	31	бы	0,46	31	вам	0,40	31	там	0,56
32	ещё	0,44	32	по	0,45	32	если	0,40	32	же	0,53
33	надо	0,44	33	надо	0,44	33	есть	0,39	33	уже	0,53
34	же	0,43	34	очень	0,43	34	то	0,39	34	надо	0,51
35	уже	0,42	35	просто	0,40	35	к	0,39	35	ты	0,51
36	есть	0,38	36	ещё	0,40	36	ты	0,39	36	можно	0,48
37	(э)	0,38	37	уже	0,38	37	или	0,38	37	но	0,47
38	тоже	0,35	38	они	0,36	38	может	0,36	38	ещё	0,44
39	нас	0,35	39	хорошо	0,35	39	будет	0,34	39	по	0,42
40	знаю	0,34	40	мы	0,34	40	дальше	0,32	40	меня	0,41
41	просто	0,34	41	будет	0,32	41	просто	0,32	41	если	0,39
42	ой	0,30	42	когда	0,31	42	хорошо	0,31	42	просто	0,38
43	тебе	0,28	43	знаю	0,31	43	уже	0,29	43	тоже	0,35
44	было	0,28	44	значит	0,30	44	потом	0,29	44	они	0,35
45	говорит	0,28	45	же	0,30	45	же	0,28	45	только	0,34
46	знаешь	0,27	46	вообще	0,29	46	(...)	0,27	46	говорю	0,32
47	очень	0,27	47	потому	0,29	47	но	0,26	47	нас	0,31
48	по	0,27	48	если	0,28	48	смотрим	0,26	48	потому	0,31
49	если	0,26	49	может	0,26	49	уровень	0,25	49	будет	0,29
50	к	0,25	50	все	0,26	50	значит	0,25	50	хорошо	0,29

Таблица 5. Наиболее употребительные слова мужской устной речи в зависимости от коммуникативной ситуации (стиля речи)

Бытовая			Профессиональная			Образовательная			«Клиент-сервис»		
Ранг	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**
1	ну	2,53	1	вот	2,77	1	это	4,15	1	вот	2,89
2	я	2,35	2	ну	2,74	2	вот	3,29	2	ну	2,54
3	не	2,32	3	да	2,57	3	я	3,16	3	в	2,18
4	в	2,12	4	я	2,49	4	да	2,44	4	я	2,01
5	вот	2,05	5	а	2,42	5	не	2,11	5	там	1,95
6	там	1,87	6	не	2,33	6	а	2,00	6	да	1,89
7	да	1,81	7	и	1,87	7	(э)	1,94	7	не	1,81
8	и	1,81	8	в	1,72	8	и	1,87	8	(э)	1,75
9	а	1,77	9	у	1,65	9	что	1,66	9	что	1,72
10	что	1,61	10	это	1,46	10	ну	1,45	10	это	1,70
11	на	1,43	11	что	1,43	11	всё	1,16	11	а	1,68
12	это	1,40	12	там	1,41	12	нет	1,12	12	у	1,56
13	у	1,12	13	так	1,32	13	на	1,10	13	и	1,56
14	так	0,97	14	на	1,13	14	так	1,05	14	так	1,43
15	ты	0,91	15	как	1,05	15	там	1,01	15	всё	1,04
16	он	0,87	16	то	1,04	16	есть	0,97	16	угу	1,04
17	(э)	0,78	17	всё	0,82	17	он	0,95	17	на	0,94
18	как	0,77	18	(э)	0,69	18	в	0,93	18	нет	0,94
19	то	0,76	19	меня	0,69	19	то	0,91	19	то	0,94
20	с	0,75	20	с	0,60	20	надо	0,84	20	с	0,89
21	всё	0,75	21	сейчас	0,58	21	здесь	0,78	21	он	0,87
22	бл**ь	0,67	22	нет	0,58	22	понятно	0,78	22	как	0,73
23	нет	0,60	23	он	0,57	23	сейчас	0,76	23	есть	0,67
24	есть	0,56	24	они	0,53	24	у	0,69	24	сейчас	0,62
25	они	0,48	25	угу	0,51	25	как	0,63	25	надо	0,54
26	меня	0,46	26	же	0,47	26	ты	0,61	26	будет	0,50
27	сейчас	0,44	27	значит	0,45	27	ага	0,59	27	мне	0,46
28	мне	0,44	28	бы	0,41	28	идёт	0,57	28	хорошо	0,46
29	просто	0,42	29	говорит	0,41	29	угу	0,51	29	если	0,46
30	уже	0,42	30	уже	0,41	30	по	0,49	30	вам	0,44
31	но	0,41	31	мне	0,41	31	она	0,49	31	они	0,44
32	ещё	0,40	32	есть	0,40	32	будет	0,44	32	уже	0,42
33	блин	0,40	33	ещё	0,40	33	значит	0,42	33	вы	0,42
34	она	0,39	34	она	0,40	34	с	0,42	34	же	0,42
35	же	0,38	35	за	0,37	35	просто	0,42	35	но	0,40
36	по	0,38	36	нас	0,37	36	уже	0,40	36	меня	0,39
37	вообще	0,37	37	блин	0,37	37	теперь	0,38	37	по	0,39
38	надо	0,35	38	ага	0,35	38	для	0,36	38	ещё	0,37
39	значит	0,34	39	просто	0,35	39	знаю	0,36	39	за	0,35
40	если	0,33	40	по	0,34	40	меня	0,34	40	ты	0,35

Бытовая			Профессиональная			Образовательная			«Клиент-сервис»		
Ранг	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**	**Ранг**	**Слово**	**%**
41	бы	0,33	41	ты	0,33	41	мне	0,34	41	значит	0,33
42	знаю	0,32	42	было	0,33	42	или	0,34	42	можно	0,33
43	короче	0,29	43	вы	0,33	43	его	0,34	43	вас	0,31
44	такой	0,28	44	тоже	0,32	44	тебе	0,34	44	или	0,31
45	мы	0,27	45	говорю	0,32	45	мы	0,32	45	может	0,29
46	потом	0,27	46	тут	0,31	46	тут	0,32	46	(м)	0,27
47	говорит	0,27	47	мы	0,30	47	но	0,32	47	семь	0,27
48	все	0,27	48	надо	0,29	48	когда	0,32	48	потом	0,25
49	тебе	0,26	49	но	0,29	49	если	0,30	49	просто	0,25
50	тут	0,25	50	если	0,29	50	самое	0,30	50	тоже	0,25

Приведенные данные показывают, что неформальная устная речь характеризуется бóльшей спецификой по сравнению с тремя другими рассматриваемыми стилями. Проявляется это, в частности, в бóльшем разнообразии в верхней зоне частотного словаря личных местоимений (особенно в женской речи), дискурсивных единиц (*ну*, *короче*, *вообще* и др.), а также в использовании в мужской речи непечатной лексики. С другой стороны, в бытовой речи обеих полов реже встречаются хезитации, «речевая поддержка» *угу*, дискурсивный маркер *вот* и такие слова, как *хорошо*, *сейчас, значит*.

6. Заключение: О состоятельности полученных списков и перспективах анализа

Поскольку сопоставимых корпусов повседневной устной русской речи не существует (что в большой степени объясняется трудоемкостью как сбора, так и обработки «живого» речевого материала), валидация полученных частотных списков в настоящее время затруднена. Для оценки их состоятельности можно предложить привлечение других, близких по объему, подвыборок рассматриваемого корпуса ОРД. Такие последовательные независимые выборки позволят оценить, при каком минимальном объеме выборочного наблюдения (количество говорящих, количество слов) в районе какого ранга наступит стабилизация верхней зоны частотного словаря [Мартыненко, 1988].

Проделанная работа носит в некоторой степени иллюстративный характер, показывая возможности исследования речевых данных с помощью специальным образом аннотированного корпуса. Тем не менее, можно полагать, что и данные списки наиболее употребительных словоформ и полученные частотные статистики, особенно в верхней зоне и для общего словаря, мужского и женского словарей в целом и их бытовой речи, на данном этапе являются удовлетворительной аппроксимацией дистрибуции наиболее употребительных словоформ, характерной для повседневной устной русской речи.

Более достоверные данные будут получены позднее в результате обработки бóльших объемов корпуса ОРД. В частности, планируется построение общего словаря повседневной речи на 1 млн словоформ, а также ряда словарей для разных социальных групп говорящих (гендерных, возрастных, профессиональных).

Литература

1. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation, Proc. 12th Int. Conf. TSD 2009, LNAI, vol. 5729, Springer, Berlin-Heidelberg, pp. 250–257.
2. *Baker P.* (2010), Sociolinguistics and Corpus Linguistics, Edinburgh University Press, Edinburgh.
3. *Bogdanova-Beglarian N., Sherstinova T., Martynenko G.* (2015), The "One Day of Speech" Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian, Proc. 18th Int. Conf "Speech and Computer" (SPECOM-2015), LNAI, vol. 9319, Springer, Switzerland, pp. 429–437.
4. *Burnard L.* (ed.) (2007), Reference guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, available at: http://www.natcorp.ox.ac.uk/docs/URG/
5. *Campbell N.* (2004), Speech & Expression; the Value of a Longitudinal Corpus, LREC 2004, Lisbon, pp. 183–186.
6. *Coupland N.* (2007), Style: Language Variation and Identity, Cambridge University Press: Cambridge.
7. *Lakoff R.* (1975), Language and Woman's Place, Harper and Row, New York.
8. *Leech G., Rayson P., Wilson A.* (2001), Word Frequencies in Written and Spoken English: based on the British National Corpus, Longman, London.
9. *Popescu, I.-I.* (2009), Quantitative Linguistics: Word Frequency Studies, Mouton de Gruyter, Berlin-New-York.
10. *Rayson P., Leech G., Hodges M.* (1997), Social differentiation in the use of english vocabulary: some analyses of the conversational component of the British National Corpus, International Journal of Corpus Linguistics, 2 (1), pp. 133–152.
11. *Romaine S.* (2008), Corpus linguistics and sociolinguistics, Corpus Linguistics: An International Handbook, Mouton de Gruyter, Berlin-New York, vol. 1, pp. 96–111.
12. *Sherstinova, T.* (2015) Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus / Ronzhin, A. et al. (eds.) SPECOM 2015, Lecture Notes in Artificial Intelligence, LNAI, vol. 9319. Springer, Switzerland, pp. 268–276
13. *Stenström, A.-B.* (1991), Expletives in the London-Lund Corpus, English Corpus Linguistics in Honour of Jan Svartvik, Longman, London, pp. 230–253.
14. *Tannen D.* (1991), You Just Don't Understand: Women and Men in Conversation. Virago Press, London.

15. *Ляшевская О. Н., Шаров С. А., Частотный словарь современного русского языка* (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. http://dict.ruslang.ru/freq.php
16. *Мартыненко Г. Я.* (1988) Основы стилеметрии. Л.: ЛГУ.
17. *Потапова Р. К., Потапов В. В.* (2006) Язык, речь, личность. М.: Языки славянской культуры. 496 с.
18. *Шайкевич А. Я.* (2015) Меры лексического сходства частотных словарей, Корпусная лингвистика — 2015. Труды международной конференции. Ответственные редакторы: Захаров В. П. , Митрофанова О. А., Хохлова М. В. С. 422–429.

References

1. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation, Proc. 12th Int. Conf. TSD 2009, LNAI, vol. 5729, Springer, Berlin-Heidelberg, pp. 250–257.
2. *Baker P.* (2010), Sociolinguistics and Corpus Linguistics, Edinburgh University Press, Edinburgh.
3. *Bogdanova-Beglarian N., Sherstinova T., Martynenko G.* (2015), The "One Day of Speech" Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian, Proc. 18th Int. Conf "Speech and Computer" (SPECOM-2015), LNAI, vol. 9319, Springer, Switzerland, pp. 429–437.
4. *Burnard L.* (ed.) (2007), Reference guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, available at: http://www.natcorp.ox.ac.uk/docs/URG/
5. *Campbell N.* (2004), Speech & Expression; the Value of a Longitudinal Corpus, LREC 2004, Lisbon, pp. 183–186.
6. *Coupland N.* (2007), Style: Language Variation and Identity, Cambridge University Press: Cambridge.
7. *Lakoff R.* (1975), Language and Woman's Place, Harper and Row, New York.
8. *Leech G., Rayson P., Wilson A.* (2001), Word Frequencies in Written and Spoken English: based on the British National Corpus, Longman, London.
9. *Lyashevskaya O. N., Sharov S. A.* (2009), Frequency List of Modern Russian language (on the Materials of the Russian National Corpus) [Chastotnyj slovar' sovremennogo russkogo yazyka (na materialah Nacional'nogo korpusa russkogo yazyka)], Azbukovnik, Moscow, available at: http://dict.ruslang.ru/freq.php
10. *Martynenko G. Ya.* (1988), Foundations of Stylometrics [Osnovy stilemetrii], Leningrad State University, Leningrad.
11. *Popescu, I.-I.* (2009), Quantitative Linguistics: Word Frequency Studies, Mouton de Gruyter, Berlin-New-York.
12. *Potapova R. K., Potapov V. V.* (2006), Language, speech, personality [Yazyk, rech', lichnost'], Yazyki slavyanskoj kul'tury, Moscow.

13. *Rayson P., Leech G., Hodges M.* (1997), Social differentiation in the use of english vocabulary: some analyses of the conversational component of the British National Corpus, International Journal of Corpus Linguistics, 2 (1), pp. 133–152.
14. *Romaine S.* (2008), Corpus linguistics and sociolinguistics, Corpus Linguistics: An International Handbook, Mouton de Gruyter, Berlin-New York, vol. 1, pp. 96–111.
15. *Shajkevich A. Ya.* (2015), Measures of lexical similarity between frequency dictionaries [Mery leksicheskogo skhodstva chastotnyh slovarey], Proc. of the Int. Conference "Corpus linguistics-2015" [Trudy mezhd. konf. "Korpusnaya linguistica-2015"], St. Petersburg State University, St. Petersburg, pp. 422–429.
16. *Sherstinova, T.* (2015), Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus, Proc. 18th Int. Conf. "Speech and Computer" (SPECOM-2015), LNAI, vol. 9319, Springer, Switzerland, pp. 268–276.
17. *Stenström, A.-B.* (1991), Expletives in the London-Lund Corpus, English Corpus Linguistics in Honour of Jan Svartvik, Longman, London, pp. 230–253.
18. *Tannen D.* (1991), You Just Don't Understand: Women and Men in Conversation. Virago Press, London.