Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2017»

Москва, 31 мая — 3 июня 2017

НЕКОТОРЫЕ ИНВАРИАНТНЫЕ ХАРАКТЕРИСТИКИ РУССКОЙ РАЗГОВОРНОЙ РЕЧИ: ФОНЕТИКА, МОРФОЛОГИЯ, СИНТАКСИС[1]

Богданова-Бегларян Н. В. (n.bogdanova@spbu.ru),
Блинова О. В. (o.blinova@spbu.ru),
Мартыненко Г. Я. (g.martynenko@spbu.ru),
Шерстинова Т. Ю. (t.sherstinova@spbu.ru)

Филологический факультет СПбГУ, Санкт-Петербург, Россия

Исследование осуществлено на материале звукового корпуса ОРД в рамках проекта, посвященного рассмотрению социолингвистической вариативности русской разговорной речи, с целью выявления диагностических признаков, характеризующих речь разных социальных групп. Результаты показали, что почти на каждом языковом уровне выявляются лингвистические параметры, в отношении которых все социолекты ведут себя одинаково: в частности, это наблюдается в дистрибуции фонем, частей речи, в частотности синтаксических структур. Распределение фонем определено на подкорпусе в 172 000 аллофонов. Наиболее частотными в речи всех социальных групп являются фонемы /a/ (18,18 %), /i/ (9,04 %), /t/ (6,36 %), /o/ (5,43 %), /u/ (4,49 %), /n/ (4,11 %), /j/ (3,82 %), /e/ (3,57 %), /k/ (3,35 %), /ы/ (3,01 %). Распределение частей речи в повседневной бытовой речи определено на лингвистически аннотированном подкорпусе объемом в 125 437 словоупотреблений и имеет следующее распределение: V (17,43 %), S (15,29 %), S-PRO (14,13 %), PART (13,35 %), CONJ (9,47 %), PR (7,09 %), ADV-PRO (5,30 %), ADV (4,51 %), A-PRO (4,30 %) и др. На синтаксическом уровне наиболее частотными в речи всех социальных групп являются одноэлементные структуры: D (3,73 %), S (2,26 %), V (1,88 %). Статистический анализ левосторонних и правосторонних зависимостей глагольных групп показал, что русский язык в его устной повседневной разновидности существенно левосторонний для всех социолектов. Представленные данные отражают некоторые постоянные, универсальные, свойства русского языка повседневного общения в целом и могут быть использованы для оптимизации систем распознавания и синтеза русской разговорной речи.

Ключевые слова: Русский язык, повседневная разговорная речь, речевой корпус, частотные списки, дистрибуция фонем, дистрибуция частей речей, синтаксические структуры

---

[1] Исследование проведено в рамках проекта «Русский язык повседневного общения: особенности функционирования в разных социальных группах», поддержанного грантом РНФ № 14-18-02070.

SOME INVARIANT FEATURES OF RUSSIAN EVERYDAY SPEECH: PHONOLOGY, MORPHOLOGY, SYNTAX

Bogdanova-Beglarian N. V. (n.bogdanova@spbu.ru),
Blinova O. V. (o.blinova@spbu.ru),
Martynenko G. Ya. (g.martynenko@spbu.ru),
Sherstinova T. Yu. (t.sherstinova@spbu.ru)

Philological Faculty of St. Petersburg State University,
St. Petersburg, Russia

The presented research was carried out on the material of the ORD speech corpus in the framework of the project, dedicated to study sociolinguistic variation of Russian speech and aimed at identifying diagnostic features characterizing everyday speech of major social groups (age-, gender-, status-, professional-related, etc.). The obtained results showed that practically on each linguistic level one may observe the features exhibiting a very high similarity between different sociolects. In particular, the coincidence is observed in the distribution of phonemes, distribution of parts of speech, and the frequency of some syntactic structures. The distribution of phonemes was determined on the subcorpus of 172,000 allophones. The following ten phonemes are the most frequent in speech of all social groups: /a/ (18,18%), /i/ (9,04%), /t/ (6,36%), /o/ (5,43%), /u/ (4,49%), /n/ (4,11%), /j/ (3,82%), /e/ (3,57%), /k/ (3,35%), /s/ (3.01%). The distribution of parts of speech in everyday speech was obtained on the linguistically annotated subcorpus of 125,437 tokens and has the following breakdown: V (17,43%), S (15,29%), S-PRO (14,13%), PART (13,35%), CONJ (9,47%), PR (7,09%), ADV-PRO (5,30 %), ADV (4,51%), A-PRO (4,30%), A (3,73%), PRAEDIC (1,84%), INTJ (1,41%), NUM (1,29%), PARENTH (0,56 %), ANUM (0,27%), PRAEDIC-PRO (0,01%). At the syntactic level, one-element structures are prevailing in everyday speech of all social groups, the most frequent among them being D (particle / discursive word) (3,73%), S (2,26%), and V (1,88%). Statistical analysis of the left-branching and right-branching verb groups has showed that the first ones significantly prevail in speech of all sociolects. The revealed features reflect some constant, universal properties of everyday spoken Russian and can be used for adjustment and improvement of speech synthesis and recognition systems.

Keywords: Russian language, everyday spoken speech, speech corpus, frequency lists, distribution of phonemes, distribution of parts of speech, syntactic structures

Введение

Корпус русской повседневной речи «Один речевой день» (ОРД) создан и продолжает пополняться в рамках большого проекта, направленного на решение ряда фундаментальных научных задач, таких как мониторинг и фиксация

звукового материала естественного языка, поддержка систем распознавания и синтеза речи, организация информационной среды и программного инструментария для нужд интегрального моделирования речи, создание источника новых учебных материалов как для изучения и преподавания языка в его звуковой форме, так и для реализации различных лексикографических и других прикладных проектов (см. подробнее: *Asinovsky et al.* 2009; *Богданова-Бегларян и др.* 2015).

На сегодняшний день корпус ОРД характеризуется следующими количественными показателями: около 1250 часов звучания, более 2800 коммуникативных макроэпизодов, 130 информантов (из них 69 мужчин и 61 женщина в возрасте от 18 до 83 лет), более 1000 основных коммуникантов. Объем текстовых расшифровок корпуса превышает 1 млн. словоупотреблений.

Основной целью, на решение которой направлено настоящее исследование, является описание современного русского языка повседневного общения и анализ особенностей его функционирования в разных социальных группах (см.: *Богданова-Бегларян и др.* 2016; *Bogdanova-Beglarian et al.* 2016 a, b). Одной из главных задач проекта является анализ материалов корпуса ОРД с целью выявления тех лингвистических параметров, по которым наиболее явно видны различия в речи говорящих, принадлежащих к различным социальным группам городского континуума. Решение данного вопроса видится как создание своеобразных *речевых портретов* городских социолектов.

Задача эта весьма нетривиальна, поскольку городская речь характеризуется несомненным *многоязычием*, а любому горожанину свойственен так называемый *полиглотизм* (термины Б. А. Ларина — *Ларин* 1977), т. е. в его речи соединяются не только черты, присущие разным социолектам (в широком смысле слова), но и разные языковые стили кодифицированного языка, на нее оказывает влияние множество внешних факторов, таких как коммуникативная ситуация, психологические характеристики всех участников разговора, их социальные роли по отношению друг к другу и собственно иерархия этих отношений, а также такие мелочи, как «внешняя обстановка: наличие или отсутствие постороннего шума, погодные условия (если разговор происходит не в помещении)» (*Фонетика спонтанной речи* 1988: 6) и др.

Несмотря на все эти допущения, мы все же предприняли попытку описать речевые портреты различных социальных групп современного города, сделать некий «*средний вывод* (курсив наш. — *Авт.*) из известного количества индивидуальных языков» (*Бодуэн де Куртенэ* 1917: 41) информантов и коммуникантов ОРД — насколько это удалось в ходе проведенного нами многоуровневого и многоаспектного анализа корпусного материала. В результате получились некоторые «штрихи к речевым портретам» разных социолектов.

Для проведения социолингвистического исследования был подготовлен расширенный исследовательский подкорпус: из общего объема транскриптов корпуса ОРД в объеме 1 млн. словоупотреблений было отобрано 100 макроэпизодов для 100 информантов, сбалансированно отражающих повседневную речь анализируемых социальных групп. Кроме речи основных 100 информантов в подкорпус была включена речь 154 коммуникантов.

В результате исследовательский подкорпус содержит речевые данные для представителей 20-ти анализируемых социальных групп: а) 2-х гендерных (мужчины и женщины), б) 3-х возрастных (молодежная, средняя и старшая), в) 10-ти профессиональных (рабочие, инженеры, военнослужащие, представители естественных наук, представители гуманитарных наук, работники образования, представители сферы обслуживания, IT-специалисты, офисные служащие, творческая интеллигенция) и г) 5-ти статусных (студенты и учащиеся; наемные работники и специалисты; руководящие работники; бизнесмены и частные предприниматели; неработающие, в том числе пенсионеры).

Объем лингвистически проаннотированного речевого материала составляет 125 437 словоформ на лексическом и морфологическом уровнях, 12 020 структур на синтаксическом уровне и 172 053 аллофонов на фонетическом уровне. Методика исследования была отработана ранее на материале пилотного подкорпуса меньшего объема (10 259 словоформ), результаты пилотного анализа представлены в коллективной монографии (*Русский язык повседневного общения*… 2016).

Результаты исследования показали, что почти на каждом лингвистическом уровне, помимо диагностических[2] и потенциально диагностических лингвистических признаков, выявляются такие, **в отношении которых все социолекты ведут себя примерно одинаково**. Эти признаки можно считать инвариантными, то есть свойственными всем говорящим по-русски и не зависящими от социальных характеристик носителя языка. В данной статье описаны такие инвариантные признаки повседневной разговорной речи, как **распределение частотных фонем**, **распределение частей речи**, а также некоторые **синтаксические особенности**, свойственные русской разговорной речи в целом.

Большинство примеров в работе извлекаются из речевого материала хорошо представленных социальных групп — гендерных (мужчины и женщины) и возрастных (младшие, средние, старшие говорящие), — проаннотированные массивы речевого материала которых в корпусе ОРД наиболее объемны.

1. Распределение частотных фонем

Статистические данные о дистрибуции фонем были получены на материале подвыборки в 172 053 аллофона, содержащей фрагменты речи всех исследуемых социальных групп.

Фонетическая транскрипция была получена автоматически посредством программного обеспечения производства ООО «Центр Речевых Технологий»[3], согласно заложенным в программе транскрибирования правилам преобразования письменного текста в последовательность аллофонов. При транскрибировании используется словарь исключений, куда экспертом вручную вносятся

---

[2] С точки зрения различения социолектов.

[3] http://www.speechpro.ru/

типичные редуцированные формы, возможные иные варианты произнесения слов, а также список типовых исключений (в частности, для разных форм редукции). Основная часть транскрипции выполняется по письменному тексту; для принятия решения о реализации слова в том случае, когда возможны варианты произнесения, программа обращается к звукозаписи, автоматически принимает решение, какой из вариантов является в данном случае наиболее вероятным, и транскрибирует слово в соответствии с этим решением.

Для транскрибирования используется следующий список аллофонов: [a0], [a1], [a2], [a4], [o0], [o1], [o4], [e0], [e1], [ы0], [ы1], [ы4], [u0], [u1], [u4], [i0], [i1], [i4], [b], [b‘], [p], [p’], [d], [d‘], [t], [t’], [g], [g‘], [k], [k’], [c], [ch], [v], [v‘], [f], [f’], [z], [z‘], [s], [s’], [zh], [sh], [sc], [h], [h‘], [m], [m’], [n], [n‘], [l], [l’], [r], [r'], [j]. Для гласных звуков 0 обозначает ударный слог, 1 — предударный, 2 — второй предударный (только для /а/), 4 — заударный.

Результаты транскрибирования и одновременного автосегментирования речевого потока на аллофоны визуализируются непосредственно в среде многоуровневого лингвистического аннотирования ELAN[4], в которой осуществляется разметка корпуса ОРД (см. рис. 1). Это позволяет провести их экспертную оценку и коррекцию по мере необходимости.

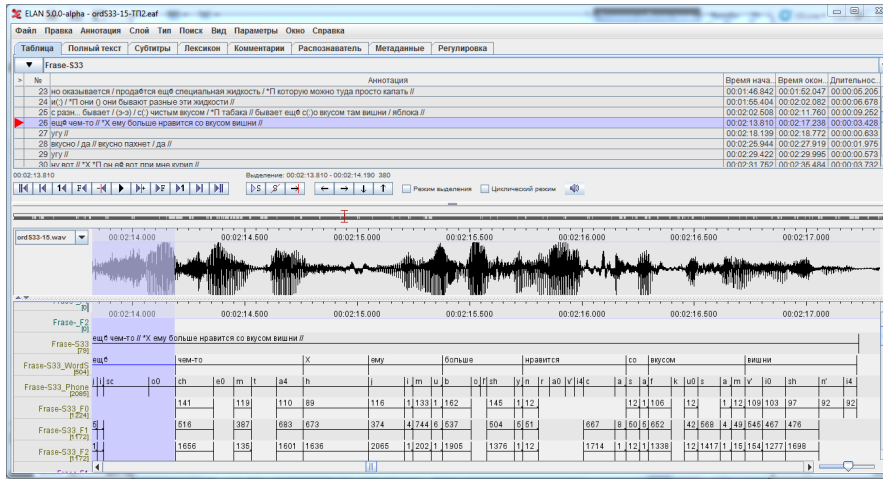

Рис. 1. Пример фонетического аннотирования корпуса ОРД (сегментация на слова и аллофоны, фонетическое транскрибирование, ЧОТ, значения первых двух формант)

Проведенное исследование показало, что наиболее частотными в речи всех социальных групп являются фонемы /а/ (18,18 %), /i/ (9,04 %), /t/ (6,36 %), /о/ (5,43 %), /u/ (4,49 %), /n/ (4,11 %), /j/ (3,82 %), /е/ (3,57 %), /k/ (3,35 %), /s/ (3,01 %). Более того, как минимум, до 10-го ранга не наблюдается абсолютно

[4] https://tla.mpi.nl/tools/tla-tools/elan/

никакого различия в рангах отдельных фонем между разными социальными группами (см. табл. 1).

Таблица 1 Верхняя зона частотного списка фонем (данные по всей выборке, мужчины и женщины, три возрастных группы)

Ранг	Всего		Гендерные группы				Возрастные группы					
	Всего	%	Муж.	%	Жен.	%	Млад.	%	Сред.	%	Стар.	%
1	**a**	18,18	**a**	17,80	**a**	18,43	**a**	18,44	**a**	17,86	**a**	17,91
2	**i**	9,04	**i**	9,02	**i**	9,05	**i**	8,76	**i**	9,21	**i**	9,70
3	**t**	6,36	**t**	6,48	**t**	6,27	**t**	6,13	**t**	6,58	**t**	6,58
4	**o**	5,43	**o**	5,61	**o**	5,31	**o**	5,26	**o**	5,60	**o**	5,48
5	**u**	4,50	**u**	4,34	**u**	4,60	**u**	4,44	**u**	4,61	**u**	4,25
6	**n**	4,11	**n**	4,15	**n**	4,08	**n**	4,33	**n**	4,01	**n**	3,94
7	**j**	3,82	**j**	3,83	**j**	3,81	**j**	3,88	**j**	3,82	**j**	3,62
8	**e**	3,57	**e**	3,50	**e**	3,61	**e**	3,62	**e**	3,46	**e**	3,62
9	**k**	3,35	**k**	3,30	**k**	3,38	**k**	3,32	**k**	3,33	**k**	3,59
10	**y**	3,01	**y**	3,05	**y**	2,99	**y**	3,04	**y**	2,96	**y**	3,15

Приведем для сравнения верхнюю зону рангового *распределения аллофонов*. Здесь для всех социальных групп на первом месте ожидаемо находится ударный [a0] (7%), на втором — [t] (6%), аллофоны [o0], [a1] и [a4] занимают позицию с третьего по пятый ранг, [n] — стабильно 6-е место в ранжированном списке, замыкают десятку аллофоны [j], [i4], [e0], [i1] и [k] (см. табл. 2).

Таблица 2. Верхняя зона частотного списка аллофонов (данные по всей выборке, мужчины и женщины, три возрастных группы)

Ранг	Всего		Гендерные группы				Возрастные группы					
	Всего	%	Муж.	%	Жен.	%	Млад.	%	Средн.	%	Старш.	%
1	**a0**	6,91	**a0**	6,81	**a0**	6,97	**a0**	7,10	**a0**	6,59	**a0**	6,73
2	**t**	6,36	**t**	6,49	**t**	6,27	**t**	6,13	**t**	6,58	**t**	6,58
3	**o0**	5,36	**o0**	5,54	**a1**	5,40	**a1**	5,29	**o0**	5,53	**o0**	5,40
4	**a1**	5,28	**a1**	5,12	**o0**	5,23	**o0**	5,19	**a1**	5,30	**a1**	5,28
5	**a4**	4,55	**a4**	4,55	**a4**	4,56	**a4**	4,61	**a4**	4,56	**a4**	4,38
6	**n**	4,11	**n**	4,15	**n**	4,08	**n**	4,33	**n**	4,01	**n**	3,94
7	**j**	3,82	**j**	3,83	**j**	3,82	**j**	3,88	**i4**	3,88	**i4**	3,91
8	**i4**	3,61	**i4**	3,73	**e0**	3,61	**e0**	3,62	**j**	3,82	**j**	3,62
9	**e0**	3,57	**e0**	3,50	**i1**	3,56	**i1**	3,62	**i1**	3,47	**e0**	3,62
10	**i1**	3,52	**i1**	3,47	**i4**	3,54	**k**	3,32	**e0**	3,46	**k**	3,59

Таким образом, в распределении отдельных аллофонов наблюдается несколько меньшая согласованность, чем в распределении частотных фонем. Однако можно предположить, что при увеличении объема исследуемой выборки ранговые последовательности аллофонов для разных социальных групп будут согласованы в большей степени.

2. Распределение частей речи

Распределение частей речи получено на лингвистически аннотированном подкорпусе ОРД. Объем исследовательского материала в словах составляет 125 437 словоупотреблений, из них 47 135 принадлежат речи мужчин, 78 302 — речи женщин. Возрастные группы представлены рабочими подкорпусами соответственно в 46 328 слов (младшая возрастная группа), 51 431 слов (средняя возрастная группа), 23 475 слов (старшая возрастная группа). Кроме того, в подкорпусе представлена речь коммуникантов-детей (4 203 слова).

Для частеречной разметки использовалась программа-морфоанализатор TreeTagger[5]. Программа использует следующий набор помет: V (глагол); S (существительное); S-PRO (местоимение-существительное); PART (частица); CONJ (союз); PR (предлог); ADV-PRO (местоимение-наречие); ADV (наречие); A-PRO (местоимение-прилагательное); A (прилагательное); PRAEDIC (предикатив); INTJ (междометие); NUM (числительное); PARENTH (вводное слово); ANUM (числительное-прилагательное); PRAEDIC-PRO (местоимение-предикатив).

Проведен автоматический морфологический анализ всего отобранного для исследования речевого материала и осуществлена его ручная коррекция. Результаты распределения частей речи в аннотированном подкорпусе представлены в табл. 3.

Таблица 3. Распределение частей речи (данные по всей выборке, мужчины и женщины, три возрастных группы)

Ранг	Всего		Гендерные группы				Возрастные группы					
	Всего	%	Муж.	%	Жен.	%	Млад.	%	Средн.	%	Старш.	%
1	V	17,43	V	17,44	V	17,42	V	17,21	V	17,65	V	17,33
2	S	15,29	S	15,68	S	15,06	S	14,86	S	15,56	S	15,30
3	S-PRO	14,13	PART*	13,40	S-PRO*	14,63	S-PRO	14,35	S-PRO	13,75	S-PRO	14,21
4	PART	13,35	S-PRO*	13,31	PART*	13,32	PART	13,80	PART	12,79	PART	13,83
5	CONJ	9,47	CONJ	9,19	CONJ	9,64	CONJ	9,31	CONJ	9,31	CONJ	10,24
6	PR	7,09	PR	7,25	PR	7,00	PR	6,99	PR	7,19	PR	7,17
7	ADV-PRO	5,30	ADV-PRO	5,68	ADV-PRO	5,08	ADV-PRO	5,32	ADV-PRO	5,52	ADV-PRO	4,88
8	ADV	4,51	ADV	4,41	ADV	4,57	ADV*	4,64	ADV*	4,53	A-PRO*	4,52
9	A-PRO	4,30	A-PRO	4,29	A-PRO	4,31	A-PRO*	4,14	A-PRO*	4,34	ADV*	4,42
10	A	3,73	A	3,92	A	3,61	A	3,72	A	3,93	A	3,41
11	PRAEDIC	1,84	PRAEDIC	1,89	PRAEDIC	1,82	PRAEDIC	1,77	PRAEDIC	1,95	PRAEDIC	1,73
12	INTJ	1,41	NUM*	1,44	INTJ*	1,55	NUM*	1,58	INTJ*	1,37	INTJ*	1,25
13	NUM	1,29	INTJ*	1,17	NUM*	1,20	INTJ*	1,47	NUM*	1,23	NUM*	0,88
14	PARENTH	0,56	PARENTH	0,60	PARENTH	0,54	PARENTH	0,56	PARENTH	0,57	PARENTH	0,61
15	ANUM	0,27	ANUM	0,31	ANUM	0,25	ANUM	0,26	ANUM	0,30	ANUM	0,20
16	PRAEDIC-PRO	0,01	PRAEDIC-PRO	0,01	PRAEDIC-PRO	0,02	PRAEDIC-PRO	0,02	PRAEDIC-PRO	0,01	PRAEDIC-PRO	0,01

[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Из табл. 3 видно, что самые распространенные части речи — это глагол (17,43 %), существительное (15,29 %), местоимение-существительное (14,13 %), частица (13,35 %) и союз (9,47 %). В целом по подкорпусу части речи имеют следующие ранги: глагол (1), существительное (2), местоимение-существительное (3), частица (4), союз (5), предлог (6), местоимение-наречие (7), наречие (8), местоимение-прилагательное (9), прилагательное (10), предикатив (11), междометие (12), числительное (13), вводное слово (14), числительное-прилагательное (15), местоимение-предикатив (16). Наблюдается почти полное совпадение ранговых последовательностей для рассмотренных 5 социальных групп.

На фоне стабильного единообразия в дистрибуции частей речи обращают на себя внимание некоторые незначительные колебания. Они относятся к гендерным группам и отмечены в таблице символом (*). Во-первых, это высокочастотная пара **S-PRO** (местоимения-существительные) и **PART** (частицы): первые чаще используют женщины, вторые — мужчины. Во-вторых, это пара менее частотных классов слов: **NUM** (числительные) и **INTJ** (междометия), имеющие 12–13 ранг (числительные более характерны для мужской речи, а междометия — для женской).

Для проверки статистической значимости различий между употреблением данных частеречных классов в речи мужчин и женщин был проведен тест с использованием стандартного критерия Стьюдента для оценки разностей выборочных долей (*Математические методы…* 1978: 181). Результаты оказались таковы:

- для частиц (PART) наблюдаемая разность между долями оказалась недостоверной при всех уровнях значимости, т. е. на данном материале нельзя утверждать, что мужчины используют частицы чаще, чем женщины;
- для местоимений-существительных (S-PRO) разность достоверна при 5%-ом уровне значимости, но недостоверна при более высоких уровнях. Поскольку в лингвистических исследованиях 5%-ый уровень значимости считается общепринятым, можно считать, что разность достоверна, т. е. женщины в повседневной речи употребляют личные местоимения чаще, чем мужчины. Эти данные согласуются и с результатами, полученными в (*Шерстинова* 2016а);
- для двух других классов слов — числительных (NUM) и междометий (INTJ) — разность между долями оказалась существенной при всех (5-, 1- и 0,1%-ом) уровнях значимости. Это означает, что мужчины действительно чаще используют в своей речи числительные, а женщины — междометия.

Ранговое распределение частей речи у представителей трёх рассматриваемых возрастных групп также практически идентично. Различия наблюдаются в рангах наречия (ADV) и местоимения-прилагательного (A-PRO): у младших говорящих и говорящих среднего возраста наречие и местоимение-прилагательное имеют 8 и 9 ранг соответственно. У старших говорящих картина обратная. При этом относительные количества употреблений упомянутых частей речи в процентах отличаются незначительно — на десятые доли процента.

Интересно, что материалы речи попавших в выборку детей (которые были собеседниками, коммуникантами информантов) и в отношении относительных цифр, отражающих употребительность слов различных частеречных классов, и в отношении их рангового распределения не показывают существенных отличий от речи взрослых говорящих.

3. Синтаксические особенности

Синтаксическое аннотирование корпуса ОРД выполнено вручную для подвыборки объемом в 13 208 элементарных предложений (клауз). Выполнялась разметка структуры глагольной группы, при этом высчитывалось количество левых и правых зависимых от вершины членов, анализировалась структура именной группы, маркировались некоторые особенности разговорного синтаксиса.

Варианты разнообразных синтаксических структур оказались весьма многочисленны, что затрудняет их формальный анализ. Наиболее частотными в речи всех социальных групп оказались одноэлементные структуры: D (частица/дискурсивное слово) (3,73 %) — *вот //, так //, да //, ладно //* и т. п., S (2,26 %) — *понедельник //, завал //, супчик?* и др., V (1,88 %) — *звони! ушел //, поехали //* и др., а также простые нераспространенные фразы структуры SV (1,07 %) — *я сваливаю //, дождь кончился //, ты сядь!* или группы дискурсивных слов {D} (1,95 %) — *вот так //, ну вот //, да да да //* и др.

Важный результат был получен при исследовании линеаризованных глагольных групп с точки зрения их левой и правой ширины (*Фитиалов* 1968) относительно вершины в структуре зависимостей. Так, каждая глагольная группа была представлена в виде кода с символом-разделителем «двоеточие», слева от которого указывается число ее левых членов (расположенных перед глаголом), а справа от него — число правых членов. При этом рассматриваются лишь количественные характеристики, при полном отвлечении от качественного состава зависимых членов в разных позициях. Ниже приведено несколько примеров таких конструкций.

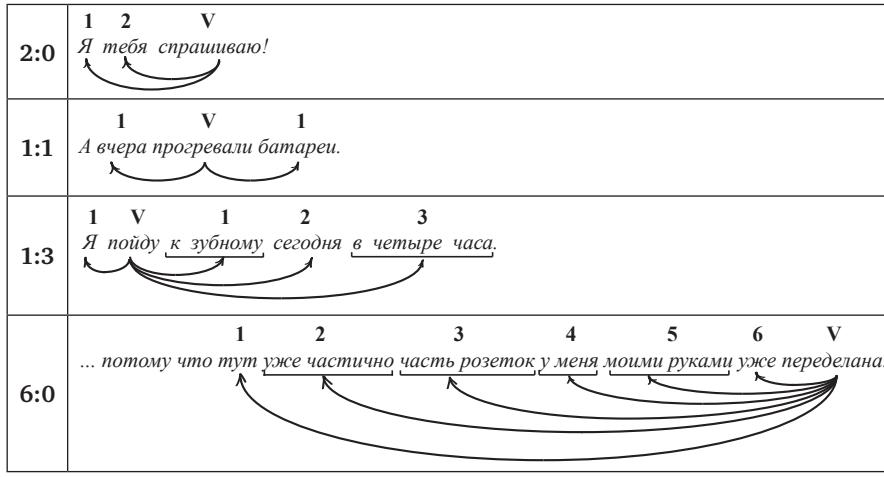

Рис. 2. Примеры линеаризованных глагольных групп с левой и правой шириной

В аннотированном подкорпусе объемом в 13 208 клауз число глагольных групп составило 7620 единиц (57,69 %). В таблице 4 приведено ранговое распределение типов глагольных групп с точки зрения их левой и правой ширины.

Таблица 4. Ранговое распределение глагольных групп с точки зрения их левой и правой ширины

Ранг	Конструкция	Кол-во	%	Ранг	Конструкция	Кол-во	%
1	1:0	1919	25,18	14	0:3	22	0,29
2	1:1	1237	16,23	15	3:2	15	0,20
3	2:0	1096	14,38	16	4:1	10	0,13
4	0:1	1084	14,23	17	0:4	8	0,11
5	0:0	729	9,57	18–19	5:0	7	0,09
6	2:1	518	6,80		2:3	7	0,09
7	3:0	352	4,62	20–21	6:0	3	0,039
8	0:2	208	2,73		1:4	3	0,039
9	1:2	168	2,21	22	3:3	2	0,026
10	3:1	98	1,29	23–25	5:1	1	0,013
11	4:0	57	0,75		5:2	1	0,013
12	2:2	51	0,67		2:4	1	0,013
13	1:3	23	0,30				

На основании этих данных могут быть построены распределения левосторонних и правосторонних членов, см. табл. 5. В таблице не принимается во внимание структура вида 0:0 (строка 5 табл. 3), так как в противном случае это приведет к искажению симметрийных отношений.

Таблица 5. Распределение левых и правых зависимых членов глагольных групп

Левая ширина	Количество	Правая ширина	Количество
0	1322	0	3434
1	3350	1	2948
2	1673	2	443
3	467	3	54
4	67	4	12
5	9	5	–
6	3	6	–
Σ	6891	Σ	6891

Оба распределения существенно разнородны с точки зрения критерия χ^2, а средние столь различны, что не нуждаются в проверке с помощью какого-либо критерия значимости. Это означает, что повседневная устная речь с точки зрения соотношения левой и правой ширины является существенно левосторонней, что подтверждает вывод, сделанный нами ранее на более ограниченном материале (*Мартыненко* 2015; *Русский язык повседневного общения*… 2016: 127–129).

Далее, с помощью критерия χ^2 было показано, что распределение левых и правых зависимостей в речи мужчин и женщин, а также в речи разных возрастных групп однородно. Кроме того, с помощью критерия Стьюдента было

установлено, что между соответствующими средними величинами в указанных группах различия статистически несущественны при самых жестких критериях значимости.

В заключение раздела приведем данные о **синтаксических нерегулярностях** повседневной речи, выделяемых на материале корпуса ОРД. При выполнении синтаксического аннотирования отмечались следующие особенности:

1) обрыв фразы (CUT):
   - *я еле люстру ...,*
   - *а утром девятого...*;
2) эллипсис (EL):
   - *он там со своим животиком //*
   - *он ещё натяжные потолки //*
   - *Оксан% / я вот вам хочу //*;
3) парцелляция (PARC), установленная на основе разметки, в которой фиксировались, в частности, длительные и синтаксически не мотивированные паузы:
   - *дырка же там вот такая // <пауза> большая //*
   - *не помню ну ладно // <пауза> потом вспомню // <пауза> скажу*
   - *тут у нас снег лежит // <пауза> прямо вообще //*;
4) самокоррекция говорящего (COR):
   - *такие глу...* (...) *ужасы рассказывают,*
   - *они* (...) *им пришлось приехать сюда.*

В табл. 5 представлены данные о распределении синтаксических особенностей устной речи, обнаруженных на материале аннотированного подкорпуса в объеме 13 020 синтаксических структур.

Таблица 5. Распределение синтаксических нерегулярностей в русской разговорной речи (данные по всей выборке, мужчины и женщины, три возрастных группы)

Ранг	Всего		Гендерные группы				Возрастные группы					
	Всего	%	Муж.	%	Жен.	%	Млад.	%	Средн.	%	Старш.	%
1	**CUT**	3,66	**CUT**	3,38	**CUT**	4,06	**CUT**	2,98	**CUT**	3,72	**CUT**	4,45
2	**EL**	1,80	**EL**	1,82	**EL**	1,79	**EL**	1,84	**EL**	1,58	**EL**	2,13
3	**COR**	1,34	**COR**	1,36	**COR**	1,32	**COR**	1,08	**COR**	1,41	**COR**	1,59
4	**PARC**	0,48	**PARC**	0,48	**PARC**	0,48	**PARC**	0,39	**PARC**	0,50	**PARC**	0,60

Из таблицы 5 видно, что в речи всех социальных групп чаще всего встречаются обрывы (незавершенные фразы), эллипсис занимает второе место по частоте встречаемости, самокоррекция — третье, реже всего используется парцелляция. Ранговые порядки отдельных синтаксических нерегулярностей инвариантны для всех рассмотренных социолектов. Просматривается общая тенденция к увеличению доли синтаксических нерегулярностей с возрастом говорящих, однако ее статистическая состоятельность требует специальной проверки.

4. Заключение

Описанные в работе данные отражают некоторые постоянные, универсальные, свойства русского языка повседневного общения в целом. Они представляют значительный теоретический интерес, а также могут быть использованы для оптимизации систем распознавания и синтеза русской разговорной речи (*Шерстинова* 2016 б). Принципиально важным представляется заключение о том, что разговорная речь, несмотря на ее спонтанность и диффузность, подчиняется достаточно жестким законам языка.

В данной работе не был рассмотрен большой пласт «потенциально инвариантных» характеристик русской разговорной речи, которые показали на проанализированном речевом материале некоторые количественные различия между рассматриваемыми группами, поскольку степень этих различий не позволяет на настоящий момент считать их статистически достоверными. К ним относится ряд характеристик лексического и лексико-дискурсивного уровней, а также другие фонетические и синтаксические особенности русской устной речи (см.: *Русский язык повседневного общения*… 2016). В отношении этих параметров необходимо проведение исследования на более представительном объеме проаннотированного речевого материала. Не исключено, что по результатам будущих исследований число инвариантных свойств русской разговорной речи будет расширено.

Литература

1. *Богданова-Бегларян Н. В., Асиновский А. С., Блинова О. В., Маркасова Е. В., Рыко А. И., Шерстинова Т. Ю.* (2015), Звуковой корпус русского языка: новая методология анализа устной речи, Язык и метод: Русский язык в лингвистических исследованиях XXI века. Вып. 2 / Ред. Д. Шумска, К. Озга. Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków, сс. 357–372.
2. *Богданова-Бегларян Н. В., Блинова О. В., Мартыненко Г. Я., Шерстинова Т. Ю.* (2016), Русский язык повседневного общения: некоторые количественные данные в зеркале соцолингвистики // Коммуникативные исследования, № 2 (8), сс. 81–92.
3. *Бодуэн де Куртенэ И. А.* (1917), Введение в языковедение. Изд. 5, Петроград.
4. *Ларин Б. А.* (1977), К лингвистической характеристике города (несколько предпосылок), Б. А. Ларин. История русского языка и общее языкознание, Просвещение, Москва, сс. 189–199.
5. *Мартыненко Г. Я.* (2015), Синтаксис живой спонтанной речи: симметрия линейных порядков, Корпусная лингвистика-2015. Труды международной конференции / Ред. В. П. Захаров, О. А. Митрофанова, М. В. Хохлова, Санкт-Петербург, сс. 371–378.
6. Математические методы в биологии. Учебное пособие (1978), под ред. Н. А. Плохинского. Изд-во МГУ, Москва.
7. Русский язык повседневного общения: особенности функционирования в разных социальных группах. Коллективная монография (2016), отв. ред. Н. В. Богданова-Бегларян. ЛАЙКА, Санкт-Петербург.

8. *Фитиалов С. Я.* (1968), Об эквивалентности грамматик НС и грамматик зависимостей, Проблемы структурной лингвистики, Наука, Москва, сс. 71–102.
9. Фонетика спонтанной речи (1988), отв. ред. Н. Д. Светозарова. Изд. ЛГУ, Ленинград.
10. *Шерстинова Т. Ю.* (2016 а), Наиболее употребительные слова повседневной русской речи (в гендерном аспекте и в зависимости от условий коммуникации), Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 1–4 июня 2016 г.), Вып. 15 (22), Изд-во РГГУ, Москва, сс. 616–631.
11. *Шерстинова Т. Ю.* (2016 б), Распознавание и синтез речи, Прикладная и компьютерная лингвистика (коллективная монография), Изд. группа УРСС, Москва, сс. 94–120.

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 / Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 57292009. Springer, Berlin-Heidelberg, pp. 250–257.
2. *Baudouin de Courtenay J.* (1917), Introduction to Linguistics [Vvedenie v yazykovedenie], Petrograd.
3. *Bogdanova-Beglarian, N. V., Asinovsky, A. S., Blinova, O. V., Markasova, E. V., Ryko, A. I., Sherstinova, T. Yu.* (2015), Speech Corpus of Russian Language: New Methodology of Speech Analysis [Zvukovoj korpus russkogo jazyka: novaja metodologija analiza ustnoj rechi], Language and Method [Jazyk i metod: Russkij jazyk v lingvisticheskih issledovanijah XXI veka] / Eds. Szumska, D., Ozga, K. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, Iss. 2, pp. 357–372.
4. *Bogdanova-Beglarian N. V., Blinova O. V., Martynenko G. Ya., Sherstinova T. Yu.* (2016), Everyday Russian Language: Some Figures in Sociolinguistic Persperctive [Russkiy yazyk povsednevnogo obshcheniya: nekotorye kolichestvennye dannye v zerkale sotsolingvistiki], Communication Studies [Kommunikativnye issledovaniya], Iss. 2 (8), pp. 81–92.
5. *Bogdanova-Beglarian, N., Martynenko, G., Sherstinova T.* (2015), The "One Day of Speech" Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian, Ronzhin, A. et al. (eds.) SPECOM 2015, Lecture Notes in Artificial Intelligence, LNAI, vol. 9319. Springer International Publishing Switzerland, pp. 429–437.
6. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Baeva E., Martynenko G., Ryko A.* (2016 a), Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech, SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811, Springer, Switzerland, pp. 659–666.
7. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G.* (2016 b), An Exploratory Study on Sociolinguistic Variation of Spoken Russian, SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 100–107.

8. Everyday Russian Language in Different Social Groups. Collective Monograph [Russkij jazyk povsednevnogo obshchenia: osobennosti funkcionirovania v raznykh social'nykh gruppakh. Kollektivnaja monografia] (2016), ed. by N. V. Bogdanova-Beglarian, LAIKA, St. Petersburg.
9. *Fitialov S. Ja.* (1968), On the Equivalence of Immediate-Constituent Grammars and Dependency Grammars [Ob ekvivalentnosti grammatik NS i grammatik zavisimostej], Problems of Structural Linguistic [Problemy strukturnoj lingvistiki], Nauka, Moscow, pp. 71–102.
10. *Larin B. A.* (1977), To the Linguistic Characteristics of the City (Several Prerequisites) [K lingvisticheskoy kharakteristike goroda (neskol'ko predposylok)], B. A. Larin, The history of the Russian language and general linguistics [Istoriya russkogo yazyka i obshchee yazykoznanie], Prosveshchenie, Moscow, pp. 189–199.
11. *Martynenko G. Ya.* (2015), Syntax of Russian Spontaneous Speech: Symmetry in Linear Order of Verb Phrase [Sintaksis zhivoi spontannoi rechi: simmetriia lineinykh poriadkov], Corpus Linguistics, Int. Conf. Proc., St. Petersburg university Publ. House, St. Petersburg, pp. 371–378.
12. Mathematical Methods in Biology [Matematicheskie metody v biologii. Uchebnoe posobie] (1978), ed. by N. A. Plokhinskiy. Moscow State University, Moscow.
13. Phonetics of Spontaneous Speech [Fonetika spontannoy rechi] (1988), ed. by N. Svetozarova, Leningrad State University, Leningrad.
14. *Sherstinova T. Yu.* (2016 a), The Most Frequent Words in Everyday Spoken Russian (in the Gender Dimension and Depending on Communication Settings) [Naibolee upotrebitel'nye slova povsednevnoy russkoy rechi (v gendernom aspekte i v zavisimosti ot usloviy kommunikatsii)], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". Issue 15 (22) [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog"]. Вып. 15 (22). RSUH, Moscow, pp. 616–631.
15. *Sherstinova T. Yu.* (2016 b), Speech Synthesis and Recognition [Raspoznavanie i sintez rechi], Applied and computational linguistics [Prikladnaya i komp'yuternaya lingvistika]. URSS, Moscow, pp. 94–120.