Санкт-Петербургский государственный университет
www.spbu.ru

НЕКОТОРЫЕ ИНВАРИАНТНЫЕ ХАРАКТЕРИСТИКИ РУССКОЙ РАЗГОВОРНОЙ РЕЧИ: ФОНЕТИКА, МОРФОЛОГИЯ, СИНТАКСИС

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)

ДИАЛОГ-2017, 02 июня 2017 г.

ОБЩАЯ ЦЕЛЬ

Проект РНФ № 14-18-02070 «Русский язык повседневного общения: особенности функционирования в разных социальных группах» (2014-2016).

Анализ особенностей функционирования современного русского языка повседневного общения в разных социальных группах.

Рассмотрение материала корпуса «Один речевой день» (ОРД), выявление лингвистических параметров, по которым различия в речи говорящих из разных соц. групп видны наиболее явно.

ОРД

Речевой корпус ОРД собирается с 2007 г. на Филологическом факультете СПбГУ.

Звукозапись речи нативных носителей русского языка, жителей СПб, в течение целого дня.

МАТЕРИАЛ

Корпус «Один речевой день» (ОРД) расширен:

- около 1250 часов звучания
- более 2800 коммуникативных макроэпизодов
- 130 информантов (69 мужчин и 61 женщина в возрасте от 18 до 83 лет)
- более 1000 основных коммуникантов
- объём текстовых расшифровок корпуса превышает 1 млн. словоупотреблений.

ИССЛЕДОВАТЕЛЬСКИЙ ПОДКОРПУС

Для проведения исследования подготовлен **исследовательский подкорпус**: отобрано **100 макроэпизодов для 100 информантов**. В подкорпус включена и речь 154 коммуникантов.

ХАРАКТЕРИСТИКА ПОДКОРПУСА

Речевые данные представителей 20-ти социальных групп, выделенных по гендерному, возрастному, профессиональному, статусному признакам.

В том числе:

2 гендерные группы (мужчины и женщины),

3 возрастные группы: младшая (18-30), средняя, (31-54), старшая, (>55 лет).

ХАРАКТЕРИСТИКА ПОДКОРПУСА

Примеры извлекаются из речевого материала хорошо представленных социальных групп – гендерных и возрастных, – проаннотированные массивы речевого материала которых в корпусе ОРД наиболее объёмны.

ОБЪЁМЫ АННОТИРОВАННОГО МАТЕРИАЛА

Объём лингвистически проаннотированного речевого материала:

125 437 словоформ на лексическом и морфологическом уровнях,

13 208 элементарных предложений (клауз) на синтаксическом уровне,

172 053 аллофона на фонетическом уровне.

Результаты исследования показали, что на разных лингвистических уровнях выявляются признаки, **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**.

КОЛОНТИТУЛ

В настоящем докладе описываются такие **инвариантные** признаки повседневной разговорной речи, как

распределение частотных фонем и аллофонов,

распределение частей речи,

а также некоторые **синтаксические особенности**, свойственные русской разговорной речи в целом.

РАСПРЕДЕЛЕНИЕ ЧАСТОТНЫХ ФОНЕМ

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

ФОНЕТИЧЕСКАЯ ТРАНСКРИПЦИЯ

Транскрипция получена автоматически с помощью ПО производства ООО «ЦРТ» согласно заложенным в программе транскрибирования правилам преобразования **письменного текста** в последовательность аллофонов.

Результаты транскрибирования и одновременного сегментирования речевого потока визуализируются в среде многоуровневого лингвистического аннотирования ELAN, в которой осуществляется разметка корпуса.

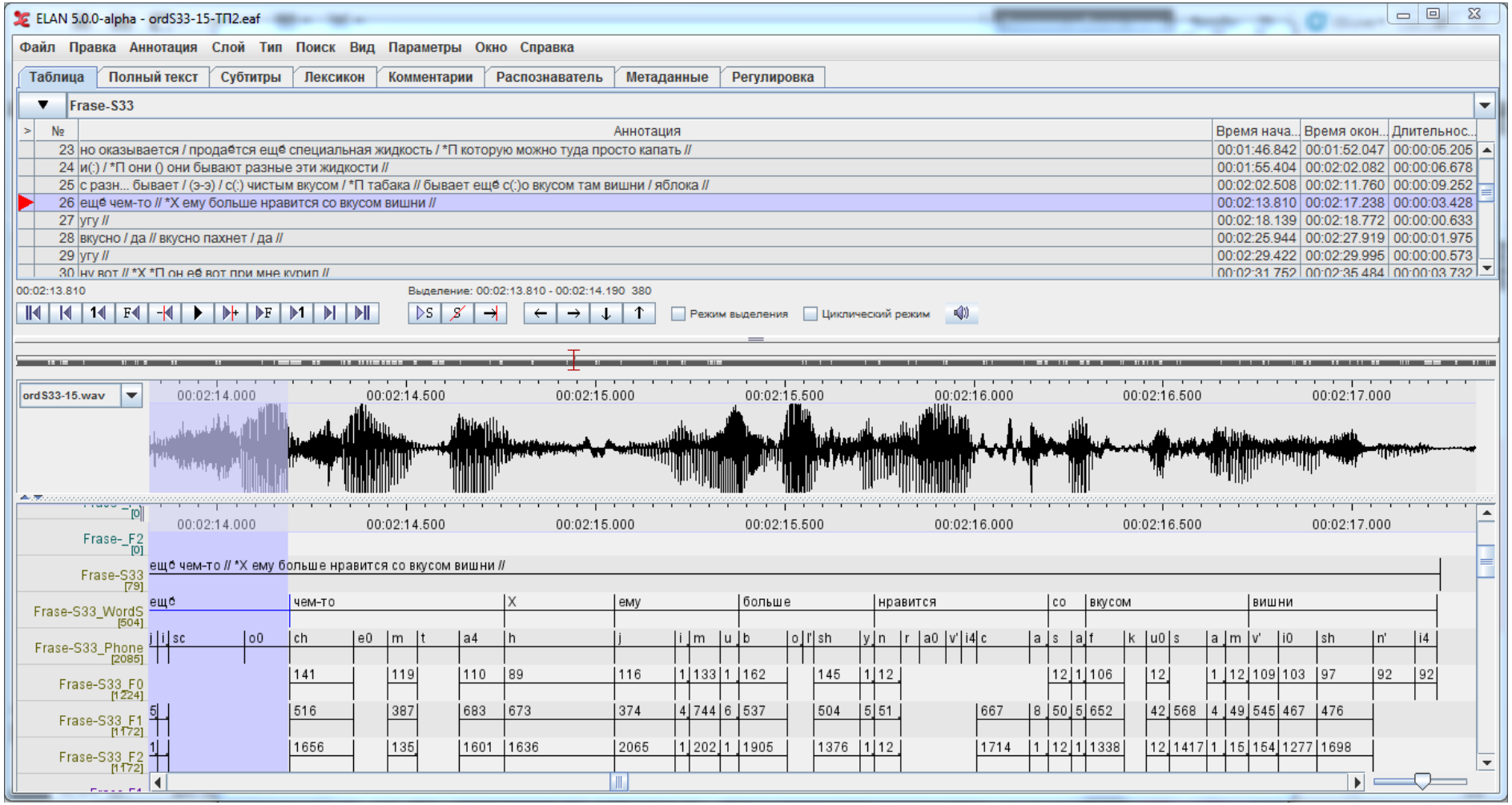
СПИСОК АЛЛОФОНОВ

При транскрибировании используется такой 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].

0 обозначает ударный слог, 1 – предударный, 2 – второй предударный (только для /а/), 4 – заударный.

ПРИМЕР ФОНЕТИЧЕСКОГО АННОТИРОВАНИЯ



РЕЗУЛЬТАТЫ (ДИСТРИБУЦИЯ ФОНЕМ)

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%), /s/ (3,01%).**

Как минимум до 10-го ранга не наблюдается абсолютно никакого различия в рангах отдельных фонем между разными социальными группами.

В распределении отдельных **аллофонов** наблюдается несколько меньшая согласованность, чем в распределении частотных фонем.

ДИСТРИБУЦИЯ АЛЛОФОНОВ

Ранг	Всего	Муж.	Жен.	Младш.	Средн.	Старш.
1	a0	a0	a0	a0	a0	a0
2	t	t	t	t	t	t
3	o0	o0	a1	a1	o0	o0
4	a1	a1	o0	o0	a1	a1
5	a4	a4	a4	a4	a4	a4
6	n	n	n	n	n	n
7	j	j	j	j	i4	i4
8	i4	i4	e0	e0	j	j
9	e0	e0	i1	i1	i1	e0
10	i1	i1	i4	k	e0	k

ДИСТРИБУЦИЯ ЧАСТЕЙ РЕЧИ

Распределение **частей речи** получено на аннотированном подкорпусе ОРД (**125 437** слов).

47 135 слов из речи мужчин,

78 302 из речи женщин.

Возрастные группы:

Младшие - 46 328 слов,

Средние - 51 431 слов,

Старшие - 23 475 слов.

Дети-коммуниканты - 4 203 слова.

ЧАСТЕРЕЧНАЯ РАЗМЕТКА

Набор частеречных 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).

ЧАСТИ РЕЧИ

Самые распространенные в подкорпусе части речи – это

глагол (17,43 %),

существительное (15,29 %),

местоимение-существительное (14,13 %),

частица (13,35 %),

союз (9,47 %).

ЧАСТИ РЕЧИ

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.

РАСПРЕДЕЛЕНИЕ ЧАСТЕЙ РЕЧИ

Ранг	Муж.	%	Жен.	%	Мл.	%	Ср.	%	Ст.	%
1	V	17,44	V	17,42	V	17,21	V	17,65	V	17,33
2	S	15,68	S	15,06	S	14,86	S	15,56	S	15,30
3	PART*	13,40	S-PRO*	14,63	S-PRO	14,35	S-PRO	13,75	S-PRO	14,21
4	S-PRO*	13,31	PART*	13,32	PART	13,80	PART	12,79	PART	13,83
5	CONJ	9,19	CONJ	9,64	CONJ	9,31	CONJ	9,31	CONJ	10,24
6	PR	7,25	PR	7,00	PR	6,99	PR	7,19	PR	7,17
7	ADV-PRO	5,68	ADV-PRO	5,08	ADV-PRO	5,32	ADV-PRO	5,52	ADV-PRO	4,88
8	ADV	4,41	ADV	4,57	ADV*	4,64	ADV*	4,53	A-PRO*	4,52
9	A-PRO	4,29	A-PRO	4,31	A-PRO*	4,14	A-PRO*	4,34	ADV*	4,42
10	A	3,92	A	3,61	A	3,72	A	3,93	A	3,41

РАСПРЕДЕЛЕНИЕ ЧАСТЕЙ РЕЧИ

Незначительные различия между группами:

S-PRO (местоимения-существительные) несколько чаще используют женщины, **PART** (частицы) – мужчины.

NUM (числительные) и **INTJ** (междометия), имеют 12–13 ранг (числительные более характерны для мужской речи, а междометия – для женской).

СТАТИСТИЧЕСКАЯ ПРОВЕРКА

Для проверки **статистической значимости** различий употребления ЧР в в речи мужчин и женщин проведен тест с использованием стандартного критерия Стьюдента для оценки разностей выборочных долей.

Для частиц 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 разность достоверна при 5 %-ом уровне значимости, но недостоверна при более высоких уровнях. Можно считать, что разность достоверна, т. е. женщины в повседневной речи употребляют местоимения-существительные чаще, чем мужчины.

Для числительных и междометий 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.

РАСПРЕДЕЛЕНИЕ ЧАСТЕЙ РЕЧИ

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.

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, маркировались некоторые особенности разговорного синтаксиса.

СИНТАКСИЧЕСКИЕ ОСОБЕННОСТИ

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 !*

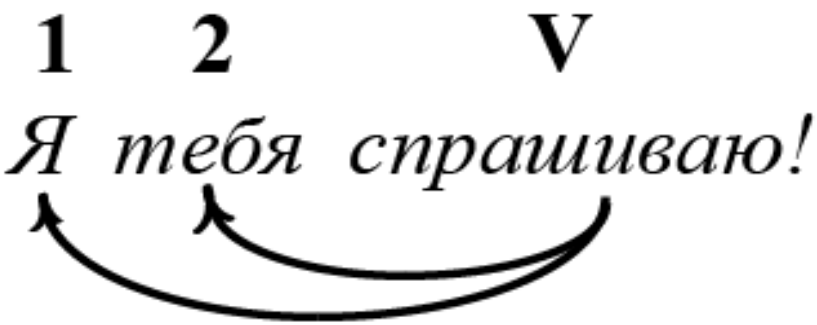
или группы частиц **{D}** (1,95 %) – *вот так //, ну вот //, да да да //* и др.

СИНТАКСИЧЕСКИЕ ОСОБЕННОСТИ

Интересные результаты были получены при исследовании линеаризованных глагольных групп с точки зрения их левой и правой ширины относительно вершины.

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 (1:0, 1:1, 2:0 и др.).

ПРИМЕР



СИНТАКСИЧЕСКИЕ ОСОБЕННОСТИ

В аннотированном подкорпусе объемом в 13208 клауз число глагольных групп составило 7620 единиц.

ЛЕВАЯ И ПРАВАЯ ШИРИНА ГЛАГОЛЬНЫХ ГРУПП

РАНГ	КОНСТРУКЦИЯ	КОЛ-ВО
1	**1:0**	1919
2	**1:1**	1237
3	**2:0**	1096
4	**0:1**	1084
5	**0:0**	729
6	**2:1**	518
7	**3:0**	352
8	**0:2**	208
9	**1:2**	168
10	**3:1**	98

РАСПРЕДЕЛЕНИЕ ЛЕВЫХ И ПРАВЫХ ЗАВИСИМЫХ ЧЛЕНОВ

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

ДАЛЬНЕЙШЕЕ ИССЛЕДОВАНИЕ

Полученные данные отражают некоторые универсальные свойства русского языка повседневного общения.

Не был рассмотрен набор **«потенциально инвариантных»** характеристик русской разговорной речи, которые на проанализированном материале показали некоторые количественные различия между рассматриваемыми группами:

- ряд характеристик лексического и лексико-дискурсивного уровней,
- некоторые фонетические и синтаксические особенности русской устной речи.

Степень этих различий, однако, не позволяет на настоящий момент считать их статистически достоверными.

ДАЛЬНЕЙШЕЕ ИССЛЕДОВАНИЕ

В отношении **«потенциально инвариантных»** параметров необходимо исследование на более представительном объеме проаннотированного речевого материала.

Не исключено, что по результатам такого исследования число инвариантных свойств будет расширено.

Санкт-Петербургский
государственный
университет
www.spbu.ru

СПАСИБО ЗА ВНИМАНИЕ!

ЛИТЕРАТУРА

- *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.
- *Baudouin de Courtenay J.* (1917), Introduction to Linguistics [Vvedenie v yazykovedenie], Petrograd.
- *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.
- *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.
- *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.
- *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.
- *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.

ЛИТЕРАТУРА

- *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.
- *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.
- *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.
- *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.
- *Mathematical Methods* in Biology [Matematicheskie metody v biologii. Uchebnoe posobie] (1978), ed. by N. A. Plokhinskiy. Moscow State University, Moscow.
- *Phonetics of Spontaneous Speech* [Fonetika spontannoy rechi] (1988), ed. by N. Svetozarova, Leningrad State University, Leningrad.
- *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.
- *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.

ДОПОЛНИТЕЛЬНЫЕ ХАРАКТЕРИСТИКИ ВЫБОРКИ

Объём исследовательского подкорпуса в словах составляет 125437 словоупотреблений.

Речевой вклад части коммуникантов (37 человек) незначительный (менее 100 слов за коммуникативный эпизод), вклад еще 76 коммуникантов умеренный (100–500 слов), остальные коммуниканты (41 человек) могут считаться полноправными "речевыми партнерами".

ДОПОЛНИТЕЛЬНЫЕ ХАРАКТЕРИСТИКИ ВЫБОРКИ

Профессиональные группы:

«гуманитарии» – 14465 слов, «естественники» – 9552 слов, «инженеры» – 8987 слов, «айтишники»

– 4048 слов, «работники образования» – 12512 слов, «офисные работники» – 15161 слов, «люди

рабочих специальностей» – 3516 слов, «силовики» – 7067 слов, «работники сферы

обслуживания» – 18744 слов, «люди творческих специальностей» – 6940 слов.

Социально-профессиональные группы:

«специалисты и наемные рабочие» – 68960 слов, «руководители в государственных и коммерческих учреждениях» – 5958 слов, «неработающие и пенсионеры» – 7635 слов, «учащиеся» – 13689 слов, «бизнесмены» – 10531 слов.

КРИТЕРИЙ СТЬЮДЕНТА

Критерий t Стьюдента направлен на оценку различий величин средних $\bar{x}$ и $\bar{y}$ двух выборок X и Y, которые распределены по нормальному закону. Одним из главных достоинств критерия является широта его применения. Он может быть использован для сопоставления средних у связных и несвязных выборок, причем выборки могут быть не равны по величине.