ЗВУКОВОЙ КОРПУС РУССКОГО ЯЗЫКА КАК БАЗА ДЛЯ СОЦИОЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ[1]

Богданова-Бегларян Н. В. (nvbogdanova_2005@mail.ru),
Шерстинова Т. Ю. (sherstinova@gmail.com),
Блинова О. В. (0973000@gmail.com),
Баева Е. М. (kaverita@yandex.ru),
Ермолова О. Б. (o-ermolova@mail.ru),
Рыко А. И. (aryko@mail.ru)

Филологический факультет СПбГУ,
Санкт-Петербург, Россия

В статье описывается методика расширения Звукового корпуса русского языка «Один речевой день» (ОРД), направленная на превращение корпуса в потенциальную базу для социолингвистических исследований. Целью таких исследований является изучение того, насколько неоднородным является наше общество с точки зрения использования бытового языка, какие группы обнаруживают наибольшее разнообразие и вариативность речевых характеристик в зависимости от конкретной коммуникативной ситуации. Для достижения этой цели прежде всего был установлен набор социальных переменных, которые учитываются при подборе информантов в мировой практике составления устных корпусов. На основе этого опыта был определен перечень социальных групп, речь которых должна войти в ОРД. После сопоставления с имеющейся базой данных корпуса выявлен перечень недостающих групп информантов и проведена дозапись материала. На данный момент ОРД содержит в целом 1017 часов звукозаписи, полученных от 107 основных информантов разных социальных и профессиональных групп (58 мужчин и 49 женщин), средний возраст которых — 37 лет (от 18 до 77 лет).

Ключевые слова: звуковой корпус, социолингвистика, русская устная спонтанная речь, социальные страты, повседневное речевое общение, городская речь, социолект

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

SPEECH CORPUS OF RUSSIAN EVERYDAY COMMUNICATION AS A RESOURCE FOR SOCIOLINGUISTIC STUDIES

Bogdanova-Beglarian N. V. (nvbogdanova_2005@mail.ru),
Sherstinova T. Y. (sherstinova@gmail.com),
Blinova O. V. (0973000@gmail.com),
Baeva E. M. (kaverita@yandex.ru),
Ermolova O. B. (o-ermolova@mail.ru),
Ryko A. I. (aryko@mail.ru)

St. Petersburg State University, St. Petersburg, Russia

The article describes methodology of expanding the ORD Speech Corpus of Russian Everyday Communication. The main aim of this expansion is to develop the database into a resource that is suitable for large-scale sociolinguistic studies. The objectives of such sociolinguistic researches are the following: 1) to reveal and describe the linguistic inhomogeneity of Russian society based on the speech, and 2) to define social groups that have the greatest variety of speech features depending on different communicative situations. To meet these goals, first of all, a set of social strata was clearly defined regarding previous scholastic experience and sociolinguistic methods used for creation of existing corpora worldwide. That determined the essential number of social groups represented in the ORD corpus. The existing recordings of the ORD corpus were analyzed based on the social group aspect with an aim to locate the gaps. Next, a list of speakers for new recordings was defined and 67 new volunteers representing different social and age strata were recorded in 2014. Now, the ORD Speech corpus of Russian Everyday Communication contains 1017 hours of recorded speech obtained from 107 informants of different social and professional groups (58 males and 49 females). The age of informants varies from 18 to 77 years with an average of 37.

Key words: speech corpus, sociolinguistics, Russian spontaneous speech, social strata, everyday speech communication, urban speech, sociolect

1. Введение

Звуковой корпус русского языка «Один речевой день» (ОРД) изначально задумывался как способ мониторинга и фиксации звукового материала естественного языка, а также организации информационной среды и программного инструментария для нужд интегрального моделирования речи (см.: *Богданова и др.* 2008; *Степанова и др.* 2009; *Asinovsky et al.* 2009 a,b). К лету 2014 г. корпус насчитывал около 500 часов звукозаписи, полученных от 40 основных информантов; выборка была сбалансирована по гендерному признаку

и содержала речь 20 мужчин и 20 женщин. Затем в течение ряда лет корпус незначительно пополнялся, а его материал аннотировался и анализировался на ряде уровней (см.: *Богданова и др.* 2010; *Звуковой корпус…* 2014, 2015).

В 2014 году перед коллективом разработчиков ОРД встала проблема существенного расширения корпуса, что было обусловлено задачами описания современного русского языка повседневного общения и анализа особенностей его функционирования в разных социальных группах (см. об этом: *Баева* 2014). Фактически была поставлена перспективная и весьма масштабная задача анализа городского «многоязычия» («многоголосия») на разных языковых уровнях, с учетом разных коммуникативных ситуаций, и решать эту задачу планировалось на материале корпуса звукозаписей повседневной коммуникации представительной выборки информантов, выполненной по методике непрерывного многочасового мониторинга речи, сбалансированно отражающих речь больших социальных групп современного мегаполиса.

Конечной ***целью*** такого направления исследований было желание понять, насколько неоднородным является наше общество с точки зрения использования бытового языка, какие группы обнаруживают наибольшее разнообразие и вариативность речевых характеристик в зависимости от конкретной коммуникативной ситуации.

Для достижения этой цели необходимо было не только расширить, но и дополнительно сбалансировать существующий корпус, что поставило перед разработчиками ряд конкретных задач: 1) выделить социальные группы, подлежащие описанию, 2) определить, какие из них и в каком объеме представлены в ОРД и, соответственно, какие группы нужно дозаписывать, 3) записать речевой материал от групп, представленных в ОРД недостаточно или не представленных вовсе. Результаты предпринятых нами шагов в решении поставленных задач и стали основой настоящей статьи.

2. О способах социолингвистической балансировки

Рассмотрим здесь коротко, какие социальные признаки в первую очередь учитывались разработчиками разных корпусов при выборе информантов. Так, в ходе формирования устной части *Британского национального корпуса* (BNC) учитывались классовый, возрастной и территориальный признаки. Отбор информантов для демографической части BNC производился согласно классификации, в рамках которой выделяются четыре *социальных класса* и шесть *возрастных групп* (см.: *Reference Guide…* 2007). При составлении *Пражского разговорного корпуса* (PMK) и *Брненского разговорного корпуса* (BMK), входящих в состав *Чешского национального корпуса*, использовалось четыре социолингвистических переменных: пол, возраст, уровень образования говорящего и «тип речи» (формальная/неформальная). Те же категории применялись при составлении «Социолингвистически сбалансированного корпуса разговорного чешского языка» (см.: *Válková, Waclawičová, Křen* 2012), кроме того, здесь учитывался регион, в котором информанты проживали в детстве.

Устный подкорпус *Национального корпуса русского языка* (НКРЯ) составлен из записей публичной и частной устной речи и транскриптов кинофильмов. Он включает тексты живой речи жителей различных регионов. Составители отмечают, что «вопрос о какой бы то ни было социологический сбалансированности устного подкорпуса НКРЯ <…> не ставится по экстралингвистическим причинам, <…> однако при конструировании устного подкорпуса мы стараемся ориентироваться, по крайней мере, на жанровую и тематическую сбалансированность» (*Гришина* 2005: 95)[2].

Более подробный обзор различных способов социолингвистической балансировки, которые использовались при составлении устных корпусов, предложен в (*Баева* 2014). На этот опыт наших предшественников и коллег мы и опирались при решении этого вопроса в ходе расширения ОРД.

3. Социальные переменные и социальные группы

Первая задача, которую мы поставили перед собой, — определение основных социальных переменных, с помощью которых предполагалось выделять группы информантов и затем балансировать выборку. Были выделены следующие социальные страты: 1) гендер, 2) возраст, 3) уровень образования, 4) род деятельности (профессия), 5) социальное положение.

Сформированный по выбранным параметрам перечень необходимых социальных групп оказался следующим.

По параметру «гендер» выделилось две традиционных категории информантов (мужчины и женщины).

По параметру «возраст» было решено выделить пять основных категорий информантов:

1) «молодежь, I группа» (18–24 года). Выбранный возрастной период — это время получения образования и начала трудовой деятельности, когда человек отрывается от родительского дома и находит свое место в социальной структуре общества;
2) «молодежь, II группа» (25–34 года). Данный возрастной период характеризуется в целом активным отношением к работе, часто — карьерным ростом, появлением собственной семьи;
3) «средний возраст, I группа» (35–44 года) — период активной работы и, как правило, расцвета социальной зрелости;
4) «средний возраст, II группа» (45–59 лет) — период от расцвета социальной зрелости до пенсионного возраста;
5) «старшее поколение» (старше 60 лет) — старшее поколение, менее активно принимающее участие в социальной жизни.

---

[2] См., тем не менее, подходы к использованию этого подкорпуса НКРЯ (правда, в оффлайновом режиме доступа) в качестве инструмента социолингвистических исследований — по крайней мере, в отношении гендера говорящих — *Даниэль, Зеленков* 2012.

По уровню образования выделяются категории информантов, получивших среднее, среднее специальное, неполное высшее, высшее образование, а также кандидаты или доктора наук.

В ходе анализа подходов к выделению групп по параметру «род деятельности» (или «профессиональная принадлежность») было решено идти двумя путями:

1) описывать речь отдельных обобщенных профессиональных групп (например, «рабочие», «работники сферы услуг» и т. п.);
2) анализировать крупные бинарные группы, противопоставленные по какому-то параметру (например, «работники физического труда» — «работники умственного труда», «гуманитарии» — «естественники», «офисные работники» — «творческая интеллигенция» и т. д.).

Принимая во внимания данные Федеральной службы городской статистики, опубликованные по результатам Всероссийской переписи населения в 2010 г. (*Социально-демографический портрет*… 2012), и особенности используемой методики звукозаписи, мы решили выбрать для анализа и сравнения 10 обобщенных профессиональных групп:

1) работники, занятые на производстве, строительстве; представители рабочих профессий, связанных с физическим трудом;
2) работники сферы услуг;
3) работники сферы образования;
4) представители силовых структур;
5) представители творческих профессий;
6) «офисные работники», специалисты по экономической деятельности, специалисты по связям с общественностью и рекламе;
7) специалисты по информационным технологиям,
8) представители инженерных специальностей;
9) представители гуманитарных наук;
10) представители естественных наук.

По социальному статусу (служебному положению) были выделены следующие категории говорящих:

1) руководящие работники и бизнесмены (руководитель, заместитель руководителя предприятия, учреждения, фирмы, департамента, отдела), частные предприниматели,
2) служащие, специалисты, наемные работники,
3) студенты и учащиеся (в том числе работающие),
4) неработающие (в том числе неработающие пенсионеры).

Получив основание для балансировки состава информантов, мы оценили, как выделенные группы соотносятся с имеющейся в ОРД базой данных.

4. Соотношение с существующей базой данных информантов ОРД

Как уже было отмечено выше, к лету 2014 г. корпус ОРД насчитывал около 500 часов звукозаписи, полученных от 40 основных информантов; выборка была сбалансирована только по гендерному признаку и содержала речь 20 мужчин и 20 женщин.

Анализ профессиональной принадлежности имеющихся информантов показал, что нет ни одной социальной группы, которая представлена в корпусе в достаточно представительном объеме. Удалось выделить следующие группы: преподаватели вузов (7 человек), инженеры (5), офисные работники (5), студенты и аспиранты (5), военнослужащие (4), работники сферы обслуживания (8), медицинские работники (3), представители гуманитарных наук (7). Остальные профессии или имеют единичных представителей, или не записаны совсем.

Таким образом, была осознана необходимость полной или частичной дозаписи представителей разных специальностей, с тем чтобы каждая из групп была представлена минимум 10-ю информантами. В первую очередь требовалась запись представителей рабочих специальностей, творческих профессий, естественных наук, специалистов по информационным технологиям, фактически не представленных в записях корпуса ОРД.

5. Методика подбора и записи информантов

Подбор добровольцев для проведения звукозаписи по методике суточного мониторинга речи велся в первую очередь с прицелом на балансировку выборки согласно выработанному списку социальных групп, таким образом, чтобы определенные для анализа социальные группы были представлены в статистически представительных объемах.

Ключевыми требованиями к кандидатурам информантов стали следующие параметры:

1) возраст (информантами могли стать люди, достигшие 18 лет),
2) родной язык (русский язык для информанта должен быть родной и основной),
3) среда проживания (город).

Существенным фактором для отбора информантов стала их личная готовность «прожить день с диктофоном на шее» и заполнить комплект социологических и психологических анкет.

Для получения новых записей был осуществлен поиск и отбор ещё 66 информантов-добровольцев. К настоящему времени дозаписано 30 женщин и 36 мужчин в возрасте от 19 до 77 лет: 36 человек в молодежной группе, 23 — в группах среднего возраста, 7 человек в группе старшего возраста.

По роду деятельности записаны следующие социальные группы (с указанием численности): работники сферы обслуживания (18), офисные работники

(13), рабочие (7), специалисты по информационным технологиям (5), инженеры (6), представители творческих специальностей (5), неработающие пенсионеры (3), преподаватели (3), представители естественных наук (2), спортсмены (3), бизнесмены (2) и один служащий силовых структур. Общая продолжительность звучания новых звукозаписей — 653 часа.

6. Адаптация информационной системы для социолингвистического расширения корпуса

Информационная система корпуса ОРД (подробнее см.: *Шерстинова* 2009) была адаптирована к задачам социолингвистического анализа речи. Так, на этапе пополнения речевого материала и формального описания макроэпизодов активно используются четыре таблицы базы данных:

1. SOC-Monitoring — мониторинг процесса звукозаписи,
2. SOC-Informants — социологическая информация об информантах,
3. SOC-Communicants — социологическая информация о коммуникантах,
4. SOC-Episodes — макроэпизоды речевой коммуникации.

Таблица SOC-Monitoring содержит следующие поля описания:

- SCode: код информанта,
- SName: псевдоним информанта,
- SGender: пол информанта,
- SAge: возраст информанта на момент записи,
- SProf: профессия или род деятельности информанта,
- Date: дата записи,
- Instructor: инструктор, ответственный за запись информанта,
- TTime: общее количество часов, записанных информантом,
- SProf: профессия или род деятельности информанта,
- Comm: комментарий по проведению записи, а также другая вспомогательная информация.

Таблица SOC-Informants содержит данные, полученные из социологической анкеты информантов, а именно:

- SCode: код информанта,
- SName: имя информанта по анкете (поскольку записи выполняются как анонимные, данное поле у большинства информантов не заполнено),
- Gender: пол информанта,
- Age: возраст информанта на момент записи,
- PBirth: место рождения,
- Gender: пол информанта,
- MLang: родной язык (на данном этапе исследования записывались исключительно информанты, для которых русский язык является родным),
- Langs: другие языки, которыми владеет информант,
- Nat: национальность родителей (по желанию),

- SClass: социальное происхождение (заполнялось в свободной форме — напр., рабочие, служащие, военные, музыканты и т. д.),
- Edu: уровень образования,
- Diploma: квалификация (специальность) информант по диплому,
- PProf: прошлые профессии или опыт работы,
- Prof: профессия или род деятельности в настоящее время,
- Regions: места длительного проживания,
- Comments: комментарии.

Для оптимизации обработки данных для некоторых параметров введены дополнительные поля, содержащие нормализованные коды. Например:

AgeGroup: возрастная группа (цифровые коды от 1 до 5);

EduN: уровень образования нормализованный (коды: В — высшее, НВ — неоконченное высшее, СС — средне-специальное, С — среднее образование);

ProfGroup: основная профессиональная группа, ProfGroupExt: профессиональная группа расширенная[3] (коды: СИЛ — представители силовых структур, ТВОР — представители творческих профессий, ИНЖ — инженеры, РАБ — рабочие, ЕСТ — ученые-естественники, ГУМ — гуманитарные специальности, ОФ — офисные работники, ОБР — работники образования, СО — работники сферы обслуживания, ИТ — IT-специалисты. Допускается одновременное использование нескольких кодов.

Status: социальное положение (коды: РУК — руководители, СН — наемные специалисты, УЧ — учащиеся, НР — неработающие).

Таблица SOC-Communicants (коммуниканты) также представляет собой данные, полученные из анкеты, заполняемой в процессе звукозаписи. Поля описания аналогичны параметрам описания информантов, но содержат два дополнительных поля: 1) SCode — код основного информанта, который обязательно указывается в дополнение к уникальному коду коммуниканта (CCode), и 2) CSRole: социальная роль коммуниканта по отношению к информанту.

Таблица SOC-Episodes содержит описание макроэпизодов речевой коммуникации. Сегментация на макроэпизоды и их аннотирование выполняется для всего речевого материала звукового корпуса (*Шерстинова* 2013). В информационную базы заносится следующая информация: SCode: код говорящего; SFName: имя звукового файла; NComType: нормализованный тип коммуникативного эпизода; NSRole: социальная роль информанта в данном эпизоде (нормализованный код); NPlace: место коммуникации (нормализованный код); SFileOrig: имя исходного (архивного) файла; Start: начальная точка эпизода относительно начала исходного файла; End: конечная точка эпизода относительно начала исходного файла; EPlace: место коммуникации (текстовое описание);

---

[3] Поскольку выяснилось, что определенный процент информантов в работает не по той специальности, по которой они получали образование, и многие информанты в настоящий могут быть отнесены к нескольким профессиональным группам (напр., преподаватель истории в вузе относится одновременно и к группе работников образования, и группе представителей гуманитарных наук), в столбце ProfGroupExt представлены все релевантные коды.

EAction: основное действие, сопутствующее разговору, или прагматическая задача; EWho: основные коммуниканты информанта в данном эпизоде; Duration: длительность эпизода (мин.); FonQuality: фонетическое качество в кодовом представлении; Priority: приоритет в расшифровке (ранговые пометы); ScenеName: содержание эпизода и комментарии; ELAN: наличие транскрипта звукозаписи (логическое поле); Comments: комментарий.

На следующих этапах реализации проекта количество таблиц базы данных будет расширено за счет введения лингвистических уровней описания: 1) фонетического, 2) лексического, 3) морфологического, 4) синтаксического.

7. Перспективы социолингвистического анализа

Нашей ближайшей задачей является определение подвыборки корпуса, на материале которой будет осуществляться сравнение лингвистических характеристик речи: в такой подвыборке для каждой исследуемой группы должны быть представлены фрагменты, отражающие речь выделенных категорий информантов в разных коммуникативных ситуациях.

Затем для отобранной подвыборки будет выполнено многоуровневое лингвистическое аннотирование и описание данных.

Так, на фонетическом уровне предполагается описать: темп речи, длительность сегментных единиц разного уровня, фонетическую реализацию частотных редуцированных форм, отдельные интонационные конструкции.

На лексическом уровне будут выделены: леммы, отмечены дискурсивные единицы; будет проанализирована функциональная активность лексических единиц (на материале частотных словарей), богатство (разнообразие) словаря, количество стилистически окрашенной, профессиональной, жаргонной и пр. лексики.

На морфологическом уровне будут размечены: части речи, грамматические признаки словоформ, аграмматичные и окказиональные формы, формы, маркированные с точки зрения нормативности.

На синтаксическом уровне будут определены: линейные структуры (порядок слов) для именных и глагольных кустов, синтаксическая сложность и распространенность речевых структур, специфические синтаксические явления устной речи (парцелляция, эллипсис, обрывы, самокоррекция и т. п.).

Обработка данных, полученных для конкретной социальной группы, и их квантитативное описание будет выполняться по мере аннотирования речевого корпуса.

8. Заключение

В результате объединения старых записей корпуса ОРД с выполненными в 2014 г. сформирован представительный массив звукозаписей в объеме более 1000 часов.

Статистические характеристики корпуса: общее количество информантов — 107 человек, из них мужчин — 58, женщин — 49. Возраст информантов — от 18 до 77 лет. По роду деятельности информанты распределяются следующим образом: работники сферы обслуживания (26), офисные работники (18), преподаватели (10), инженеры (11), представители творческих специальностей (7), рабочие (7), специалисты по информационным технологиям (6), служащие силовых структур (5), представители гуманитарных наук (9), неработающие пенсионеры (3), представители естественных наук (3), спортсмены (3). Пополнение корпуса новыми звукозаписями будет продолжено, с тем чтобы каждая описываемая социальная группа была представлена как минимум 10 информантами.

Текстовые расшифровки корпуса ОРД насчитывают порядка 500 000 словоупотреблений. Планируется довести этот объем до 1 млн Продолжается аннотирование звукового материала по 7-ми стандартным уровням, принятым в корпусе ОРД (обязательные уровни: реплика с указанием синтагматического членения, код говорящего, коммуникативный микроэпизод; факультативные параметры: невербальные аудиособытия, качество голоса, фонетический комментарий, общий комментарий).

На материале корпуса ОРД выполняются пилотные социолингвистические исследования. В первую очередь следует отметить исследования темпа речи. Так, была обнаружена значимая зависимость этого параметра от социальных факторов — гендера (мужчины говорят быстрее женщин: 5,5 vs. 5,3 сл./с) и возраста (с возрастом мы говорим медленнее: 6,1 vs. 5,3 сл./с). Кроме того, оказалось, что петербуржцы (среди информантов ОРД представлены не только коренные петербуржцы) говорят медленнее, чем приезжие из других городов (5,1 vs. 5,78 сл./с) (см. об этом подробнее: *Stepanova* 2013). Была выявлена и определенная зависимость темпа речи от социальной роли информанта и, соответственно, от коммуникативной ситуации. Например, информант S40 (40 лет, врач-педиатр) в процессе общения с матерью пациента говорит намного медленнее, чем в среднем в течение своего речевого дня (при общении с коллегами или женой) (*Метлова* 2014). Эти предварительные наблюдения будут проверены на представительном корпусном материале, а также дополнены социолингвистическими исследованиями в области фонетики, лексики, морфологии и синтаксиса.

В заключение следует отметить, что осуществляемое расширение корпуса ОРД будет способствовать не только проведению на его базе социолингвистических исследований разного типа, но может найти свое применение и для многоаспектного описания русского устного дискурса, прагматики, грамматики речи, когнитивной лингвистики, психолингвистики, криминальной фонетики, а также решения задач современных речевых технологий, связанных с синтезом и распознаванием естественной речи.

Литература

1. *Баева Е. М.* (2014), О способах социолингвистической балансировки устного корпуса (на примере «Одного речевого дня»), Вестник Пермского университета. Российская и зарубежная филология, Вып. 4 (28), с. 48–57.
2. *Богданова Н. В., Асиновский А. С., Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.* (2009), Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка, Компьютерная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог'2009», Москва, РГГУ, с. 38–44.
3. *Богданова Н. В., Асиновский А. С., Маркасова Е. В., Степанова С. Б., Супрунова А. В., Шерстинова Т. Ю.* (2010), Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования, Компьютерная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог'2010», Москва, РГГУ, с. 41–46.
4. *Гришина Е. А.* (2005), Устная речь в Национальном корпусе русского языка, Национальный корпус русского языка: 2003–2005, Москва, Индрик, с. 94–110.
5. *Даниэль М. А., Зеленков Ю. Г.* (2012), НКРЯ как инструмент социолингвистических исследований. Эпизод IV. пол говорящего и длина реплики, Компьютерная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог'2012», Москва, РГГУ, с. 112–121.
6. *Ерофеева Т. И.* (2009), Социолект: Стратификационное исследование, Пермь, Перм. гос. ун-т.
7. Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 2. Теоретические и практические аспекты анализа. Том 1. О некоторых особенностях устной спонтанной речи разного типа. Звуковой корпус как материал для преподавания русского языка в иностранной аудитории (2014), Отв. ред. Н. В. Богданова-Бегларян, Санкт-Петербург, Филологический факультет СПбГУ.
8. Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 2. Теоретические и практические аспекты анализа. Том 2. Звуковой корпус как материал для новых лексикографических проектов (2015), Отв. ред. Н. В. Богданова-Бегларян, Санкт-Петербург, Филологический факультет СПбГУ.
9. *Метлова В. А.* (2014), Темп речи в свободной коммуникации: социолингвистический аспект, Вестник Пермского университета. Российская и зарубежная филология, вып. 3 (27), с. 58–65.
10. *Социально-демографический портрет* (2012), Социально-демографический портрет России. По итогам Всероссийской переписи населения 2010 года, Федер. служба гос. Статистики, Москва, ИИЦ «Статистика России».
11. *Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю.* (2008), Звуковой корпус русского языка повседневного общения «Один речевой день»: Концепция и состояние формирования, Компьютерная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог'2008», Москва, РГГУ, с. 488–494.

12. *Шерстинова Т. Ю.* (2013), Коммуникативные макроэпизоды в корпусе повседневной русской речи «ОДИН РЕЧЕВОЙ ДЕНЬ»: принципы аннотирования и результаты статистической обработки, Труды Международной конференции «Корпусная лингвистика-2013», Санкт-Петербург, Филологический факультет СПбГУ, с. 449–456.

References

1. *Baeva E. M.* (2014), On Means of Sociolingiustic Balancing of a Spoken Corpus (Drawn on one Speaker's Day) [O sposobax sociolingvisticheskoj balansirovki ustnogo korpusa (na primere "Odnogo rechevogo dn'a")], Perm University Herald. Russian and Foreign Philology [Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia], Iss. 4 (28), pp. 48–57.
2. *Bogdanova, N. V., Asinovsky, A. S., Rusakova, M. V., Ryko, A. I., Stepanova, S. B., Sherstinova, T. Ju.* (2009), Speech Corpus as a Tool for Monitoring and Fixation of Various Forms of Natural Language [Zvukovoj korpus kak sposob monitoringa i fiksacii raznyx form estestvennogo jazyka], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2009"], Moscow, pp. 38–44.
3. *Bogdanova N. V., Asinovsky A. S., Markasova E. V., Stepanova S. B., Suprunova A. V., Sherstinova T. Ju.* (2010), The ORD Speech Corpus of Russian Everyday Communication: Ways of Replenishment and First Results of Analisys [Zvukovoj korpus russkogo jazyka "Odin rechevoj den": puti popolnenija i pervyje rezultaty issledovanija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2010"], Moscow, pp. 41–46.
4. *Grishina E. A.* (2005), Spoken Speech in National Russian Corpus [Ustnaja rech' v Nacional'nom korpuse russkogo jazyka], National Russian Corpus 2003–2005 [Nacional'nyj korpus russkogo jazyka: 2003–2005], Moscow, pp. 94–110.
5. *Daniel M. A., Zelenkov Yu. G.* (2012), Russian National Corpus as a Playground for Sociolinguistic Research. Episode IV. Gender and the Length of the Utterance [NKRJa kak instrument sociolingvisticheskix issledovanij: epizod IV. Pol govor'ashchego i dlina repliki], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2012"], Moscow, pp. 112–121.
6. *Erofeeva T. I.* (2009), Sociolect: Stratification Research [Sociolekt: Stratifikacionnoe issledovanie], Perm.
7. *Metlova V. A.* (2014), Speech tempo in free communication: a sociolinguistic aspect [Temp rechi v svobodnoj kommunikacii: sociolingvisticheskij aspekt], Perm University Herald. Russian and Foreign Philology [Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia], iss. 3 (27), pp. 58–65.

8. Speech Corpus as a Base for Analysis. Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 1. Some Features of Various Species of Oral Spontaneous Speech. Speech Corpus as a Material for Teaching Russian for Foreigners (2014), [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografija. Chast' 2. Teoretichskie i prakticheskie aspekty analiza. Tom 1. O nekotoryx osobennost'ax ustnoj rechi raznogo tipa. Zvukovoj korpus kak material dl'a prepodavania russkogo jazyka v inoctrannoj auditorii], Sankt Petersburg.
9. Speech Corpus as a Base for Analysis. Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 2. Speech Corpus as a Base for New Lexicographical Projects (2015), [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografija. Chast' 2. Teoretichskie i prakticheskie aspekty analiza. Tom 2. Zvukovoj korpus kak material dl'a novyx leksikograficheskix proektov], Sankt Petersburg.
10. *Social and Demographic Portrait of Russia: the Result of Population Census of 2010* (2012), [Social'no-demograficheskij portret Rossii: Po itogam Vserossijskoj perepisi naselenija 2010 goda], Russian Federal State Statistics Service [Federalnaja sluzhba gosudarstvennoj statistiki], Moscow.
11. *Stepanova S. B., Asinovsky A. S., Bogdanova N. V., Rusakova M. V., Sherstinova T. Ju.* (2008), Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Basic Conception and Current State [Zvukovoj korpus russkogo jazyka povsednevnogo obschenia: Kontseptsia i sostojanie formatirovania], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2008" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2008"], Bekasovo, pp. 488–494.
12. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), Speech Corpus of Russian Everyday Communication «One Speaker's Day» (the ORD corpus), SPECOM'2009. Proceedings of the 13th International Conference «Speech and Computer» (2125 June 2009, St. Petersburg, Russia), Sankt Petersburg, pp. 521–526.
13. *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, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 5729 LNAI, pp. 250–257.
14. Reference Guide for the British National Corpus [XML Edition]. Ed. by L. Burnard, publ. for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services, 2007, available at: http://www.natcorp.ox.ac.uk/docs/URG.xml.
15. *Sherstinova T.* (2009), The Structure of the ORD Speech Corpus of Russian Everyday Communication, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 5729 LNAI, pp. 258–265.
16. *Sherstinova T.* (2013), Communicative Macro Episodes in the ORD Corpus of Russian Everyday Communication: Annotation Principles and Some Statistics [Kommunicativnye macroepisody v korpuse povsednevnoj russkoj rechi "Odin

rechevoy den'": principy annotirovaniya i rezul'taty statisticheskoy obrabotki], Proc. of the International Conference "Corpus linguistics-2013" [Trudy mezhdunarodnoy nauchnoy konferentsii "Korpusnaya linguistica-2013"], St. Petersburg, pp. 449–456.

17. *Stepanova S.* (2013), Speech Rate as Reflection of Speaker's Social Characteristics, Approaches to Slavic Interaction, Amsterdam: John Benjamins Publishing Company, 2013, pp. 117–129.
18. *Válková L., Waclawičová M., Křen M.* (2012), Balanced data repository of spontaneous spoken Czech // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) / İstanbul: ELRA, pp. 3345–3349, available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/179_Paper.pdf.