Н. В. Богданова-Бегларян, О. В. Блинова, К. Д. Зайдес, Т. Ю. Шерстинова
Санкт-Петербургский государственный университет
(Россия, Санкт-Петербург)
n.bogdanova@spbu.ru, o.blinova@spbu.ru,
kristina.zaides@student.spbu.ru, t.sherstinova@spbu.ru

КОРПУС «СБАЛАНСИРОВАННАЯ АННОТИРОВАННАЯ ТЕКСТОТЕКА» (САТ): ИЗУЧЕНИЕ СПЕЦИФИКИ РУССКОЙ МОНОЛОГИЧЕСКОЙ РЕЧИ*

Статья представляет один из корпусов русской устной речи: коллекцию спонтанных монологических текстов, известную как «Сбалансированная аннотированная текстотека» (САТ). Данный корпус собирается в Санкт-Петербургском государственном университете в течение уже более чем 20 лет с использованием авторской (Н. В. Богдановой-Бегларян) методики сбора данных, предполагающей достаточно строгий набор экспериментальных процедур. САТ предназначендля изучения спонтанных монологов разного типа (чтение (сюжетного и несюжетного исходных текстов), пересказ прочитанных текстов, описание изображения (также сюжетного и несюжетного), рассказ на заданную тему) и содержит тексты, записанные от пяти профессионально-ориентированных групп носителей языка (медики; юристы; «компьютерщики»; филологи, преподаватели русского языка как иностранного; преподаватели-философы), несколько блоков речи студентов (филологов и нефилологов), а также четыре блока интерферированной русской речи носителей других языков: американского английского, китайского, французского и нидерландского. Всего в составе САТ сегодня около 700 текстов и около 50 часов звучания. В статье на фоне других русскоязычных и иноязычных устных корпусов дано описание данного лингвистического ресурса, отмечены основные темы, разрабатываемые на его материале, а также намечены перспективы продолжения работы.

Ключевые слова: современный русский язык, устная монологическая речь, звуковой корпус, обработка естественного языка, база данных, лингвистический

* Исследование выполнено при поддержке гранта РФФИ № 17-29-09175 «Диагностические признаки социолингвистической вариативности повседневной русской речи (на материале звукового корпуса)».

эксперимент, чтение, описание изображения, пересказ текста (репродуктив), спонтанный монолог, социолингвистика, психолингвистика.

Исследованием разных аспектов русской спонтанной монологической речи в разные годы занимались многие лингвисты, что лишний раз свидетельствует о важности этой темы для традиционной русистики. В этой связи следует отметить работы фонетистов — Л. В. Бондарко, Н. Д. Светозаровой, Н. Б. Вольской, Н. И. Гейльман, О. Ф. Кривновой, С. В. Кодзасова, Л. Л. Касаткина, психолингвистов — Л. С. Выготского, Н. И. Жинкина, А. А. Леонтьева, А. Р. Лурии, И. Н. Горелова, К. Ф. Седова, социолингвистов — Б. Н. Головина, Л. П. Крысина, А. П. Мартынюка, Т. И. и Е. В. Ерофеевых, А. В. Кирилиной, М. Краузе, специалистов в области лингвистической экспертизы — к примеру, Е. И. Галяшиной, диалектологов — Е. А. Оглезневой, И. А. Букринской, О. Е. Кармаковой. В контексте исследования устного монолога необходимо упомянуть также имена Н. Ю. Шведовой, О. А. Лаптевой, О. Б. Сиротининой, Е. В. Красильниковой и ряд других. Однако эта ниша в коллоквиалистике все еще нуждается в детальном и многоаспектном исследовании. Именно этой цели служат материалы корпуса «Сбалансированная аннотированная текстотека» (САТ), которому и посвящена настоящая статья.

САТ представляет собой коллекцию монологических текстов, которая собирается в Санкт-Петербургском государственном университете в течение уже более чем 20 лет (см., например: [Богданова и др. 2008; Богданова 2010; Богданова-Бегларян и др. 2017а]) с использованием авторской (Н. В. Богдановой-Бегларян) методики сбора данных, предполагающей достаточно строгий набор экспериментальных процедур. САТ — это во многом уникальный лингвистический ресурс, предназначенный для изучения спонтанных монологов разного типа.К настоящему времени корпус содержит тексты, записанные от пяти профессионально-ориентированных групп носителей языка (медики; юристы; «компьютерщики»; филологи, преподаватели русского языка как иностранного и преподаватели-философы), несколько блоков речи студентов (филологов и нефилологов), а также четыре блока интерферированной русской речи носителей других языков: американцев, китайцев, франкофонов и голландцев (см. подробнее об исследованиях на материале русской речи иностранцев: [Метлова 2013; Казак 2015; Зайдес 2016, 2017; Замковец 2018; Чэн Чэнь 2018]). Всего в составе САТ сегодня около 700 текстов и около 50 часов звучания.

С учетом состава информантов монологи в составе САТ можно разделить на следующие группы:

- 150 монологов медиков;
- 201 монолог юристов;
- 172 монолога студентов (филологи и нефилологи);
- 32 монолога преподавателей РКИ;
- 28 монологов «компьютерщиков»;
- 12 монологов преподавателей-философов;
- 12 монологов информантов разных профессий;

• 62 монолога изучающих русский язык инофонов (8 текстов, записанных от носителей американского варианта английского языка, 8 — от франкофонов, 30 — от носителей китайского языка, 16 — от носителей нидерландского языка).

Данный материал по типам спонтанных монологов распределяется следующим образом:

• чтение предтекста — 72 монолога (36 монологов-чтений сюжетного и 36 — несюжетного предтекстов);

• пересказ первичного текста — 204 монолога (101 пересказ сюжетного предтекста и 103 — несюжетного);

• описание изображения — 254 монолога (136 описаний сюжетного изображения и 118 — несюжетного);

• рассказ — 139 монологов.

Общее количество информантов, речь которых вошла в САТ, — 212 человек. Состав информантов с учетом их социальных и психологических характеристик:

1) по гендерному признаку: 89 мужчин, 123 женщины;

2) по возрасту:

• 1 группа, поздняя юность (18–24 года) — 85 информантов;

• 2 группа, ранняя зрелость (25–34 года) — 41 информант;

• 3 группа, средняя зрелость (35–44 года) — 39 информантов;

• 4 группа, поздняя зрелость (45–54 года) — 13 информантов;

3) по уровню речевой компетенции:

• 43 информанта с высоким УРК;

• 54 информанта со средним УРК;

• 73 информанта с низким УРК;

4) по профессиональной принадлежности:

• 95 студентов (носителей русского языка и иностранцев);

• 48 юристов;

• 30 медиков;

• 7 преподавателей РКИ;

• 14 «компьютерщиков»;

• 6 преподавателей-философов;

• 12 информантов разных профессий;

5) по психологическим характеристикам (представлены не для всех групп информантов; общее количество говорящих, психологические характеристики которых известны, — 97 человек):

• 38 экстравертов;

• 31 амбиверт;

• 28 интровертов.

Тексты для корпуса записывались в разное время, разными собирателями и часто в разных целях (как правило, в рамках той или иной исследовательской работы, от студенческой курсовой до кандидатской диссертации), первые записи были сделаны еще до рождения самой идеи корпуса [Степихов 2005; Бродт 2007], поэтому материал САТ не вполне однороден и далеко не все его блоки отвечают тем

требованиям, которые позже были сформулированы. Требования эти (принципы создания корпуса) в общем и целом таковы.

Коллекция САТ предполагает *балансировку* и материала, и состава информантов:

• *собственно лингвистическая балансировка*: все тексты записывались по специально разработанной программе, в рамках разных коммуникативных (речевых) сценариев: чтение и пересказ сюжетного и несюжетного текстов, описание сюжетного и несюжетного изображений, рассказ на заданную тему; для некоторых конкретных исследований записывались тексты только одного типа, которые и подвергались углубленному анализу: *чтение* [Сапунова 2009], *пересказ* [Куканова 2009; Малиновская 2012], *описание* [Филиппова 2010], *рассказ* [Иванова 2011; Казак 2015; Зайдес 2016];

• *социолингвистическая балансировка*: информанты подбираются с учетом их социальных характеристик, что фиксируется в специальных анкетах и затем попадает в базу данных корпуса;

• *психолингвистическая балансировка*: информанты обладают разными психологическими характеристиками (психотип и в ряде исследований — уровень невротизма/нейротизма), что устанавливалось в ходе специального психологического тестирования (тест Г. Айзенка).

Лингвистическая (подбор текстов-стимулов) и социолингвистическая (подбор информантов) балансировка проводились до записи, психологическое тестирование участников эксперимента — после записи. В результате полной психолингвистической балансировки не получилось, но в целом материал САТ дает некоторую возможность его анализа с учетом психологических факторов (см., например: [Куканова 2009; Хан 2013; Зайдес 2016]).

За годы существования САТ был проанализирован и продолжает исследоваться в самых разных направлениях. Можно сказать, что на его основе проводится фундаментальное многоуровневое (фонетика, грамматика, лексика, структура дискурса) описание звучащей монологической речи, которая является основной формой человеческого общения. Фундаментальность исследований, проводимых на материале САТ, определяется до некоторой степени новым подходом к анализу монологической речи.

Прежде всего в научный оборот вводится ряд новых понятий: *профессиональное/непрофессиональное отношение* говорящего *к языку/речи*, *степень естественности* и *спонтанности* устной речи, степень *лингвистической мотивированности* устного монологического текста, *уровень речевой компетенции* (УРК) говорящего. В целом ряде работ авторского коллектива предлагается исследование этих понятий в широком междисциплинарном контексте (с привлечением методов психолингвистики, социолингвистики, дискурсивного анализа, акустического инструментального анализа). Пилотный анализ показал, что эти параметры действительно являются дистинктивно значимыми для анализа спонтанных монологов разных типов (см., например, трехтомную коллективную монографию: [Звуковой корпус… 2013, 2014, 2015]). При аннотировании корпуса САТ каждый звучащий текст был атрибутирован с точки зрения степени его спонтанности

и лингвистической мотивированности. База данных CAT содержит информацию обо всех информантах, участвовавших в записи.

Другая важная научная задача, решению которой служит корпус CAT, состоит в исследовании степени вариативности порождаемых монологических текстов в зависимости от характера стимула: визуального или текстового. В частности, предварительные наблюдения показывают, что сюжетность/несюжетность предтекста или описываемого изображения, а также степень знакомства говорящего с темой свободного монолога, заданного вопросом экспериментатора (исходным стимулом), оказывают влияние на выбор говорящим тех или иных речевых средств и в целом на лингвистическую природу вторичного текста (см. многочисленные наблюдения в: [Звуковой корпус… 2013, 2014, 2015]). Этот аспект порождения устных монологических текстов еще мало изучен.

Социолингвистическая и психолингвистическая вариативность речевой продукции в жанре монолога — еще один важный аспект исследования корпуса CAT. Предлагаемая разработчиками CAT методика позволяет анализировать речевую продукцию говорящих с разными психологическими и социологическими характеристиками, полученную фактически в одинаковых коммуникативных ситуациях, что дает основания для ее научного анализа в сопоставительном аспекте.

Спонтанная устная речь представляет собой весьма сложный для научного анализа объект. По форме выражения — по крайней мере, внешне — она значительно менее упорядочена и структурирована, чем письменная, выглядит более хаотичной и непредсказуемой, что, тем не менее, не мешает ей выполнять свои основные функции. Поэтому в настоящее время лингвисты все чаще говорят о принципиальном отличии *грамматики* спонтанной устной *речи* от существующих академических грамматик, описывающих стандартный, кодифицированный, литературный язык в его письменной форме. Правила порождения спонтанной речи и ее грамматика до сих пор системно не описаны, приблизиться к решению этой задачи может помочь анализ материала CAT.

Более того, наблюдения показывают, что не только грамматика речи обладает существенным своеобразием по сравнению с письменными формами языка, но и *грамматика* устного *монолога* в определенных аспектах принципиально отличается от *грамматики* устного *диалога* на фонетическом, морфологическом, синтаксическом, лексическом и дискурсивном уровнях (см., например: [Богданова 2012; Завадская 2018]). В первую очередь эти различия касаются принципов текстообразования монологической спонтанной речи и описательного дискурса, которые непосредственно влияют на синтаксическую организацию текста. Такой аспект анализа становится возможен при сравнении материала CAT (монологические тексты) и корпуса повседневной русской речи «Один речевой день» (ОРД; в основном диалоги и полилоги) (см. о нем, например: [Asinovsky et al. 2009; Русский язык повседневного общения… 2016; Bogdanova-Beglarian et al. 2016a, b, 2017; Богданова-Бегларян и др. 2017б]).

Материалы CAT становятся доступными широкому кругу читателей и пользователей двумя путями.

Первый — через серию публикаций текстов устной спонтанной речи, которая начала выходить с 2008 г.: в первых трех выпусках этой серии содержатся спонтанные тексты, записанные от информантов-юристов, — свободные рассказы [Русская спонтанная речь 2008], монологи-репродуктивы [Русская спонтанная речь 2010] и монологи-описания [Русская спонтанная речь 2011]. Сборник текстов, записанных от информантов-медиков (монологи всех типов), готовится к печати в Германии [Русская спонтанная речь 2020]. Отличительной чертой этих публикаций является наличие, помимо, собственно транскриптов всех текстов и метаданных информантов, лексических материалов: частотных списков, как общих, так и упорядоченных по типам текстов.

Второй путь, которым материалы САТ могут попасть в руки широкого круга читателей и пользователей, — их постепенная передача в Национальный корпус русского языка (его устный подкорпус). Эта работа проводится в настоящее время. Однако предпринимаемых усилий по «сближению» материалов САТ с их потенциальным пользователем пока явно недостаточно. Поэтому в перспективах работы авторского коллектива — разработка системы онлайн-доступа к звуковому корпусу русской монологической речи.

Возвращаясь к существующим исследованиям спонтанной речи, еще раз повторим, что центральным объектом внимания лингвистов по сей день остается русская разговорная речь в форме бытового диалога: в работах исследователей оказывается затронутой специфика спонтанной речи, которая раскрывается прежде всего на материале записанных и расшифрованных диалогов. Бытовые спонтанные монологи как исследовательский материал представлены явно недостаточно.

Традиционно для сбора речевых данных используется ряд методик. Наиболее распространенными из них являются устное интервью, разнообразные эксперименты, включенное наблюдение.

Из отечественных работ следует отметить, например, записи, сделанные М. В. Китайгородской и Н. Н. Розановой и нашедшие отражение в книгах «Речь москвичей: коммуникативно-культурологический аспект» (1999) и «Языковое существование современного горожанина: На материале языка Москвы» (2010). В сборнике «Жанр интервью: Особенности русской устной речи в Финляндии и Санкт-Петербурге» (2004) анализировался корпус текстов полуструктурированных интервью, записанных российско-финским коллективом социологов (И. И. Травин, Е. М. Порецкина, Т. Пиирайнен, Ю. Симпура и др.) от жителей Санкт-Петербурга в 1999–2003 гг., а также тексты, записанные от русскоговорящих жителей Финляндии. Необходимо упомянуть также о записях русской речи коми-пермяков и татар, собранных под руководством Т. И. Ерофеевой (2007, 2010, 2012). Материалы русской речевой коммуникации (монологи, диалоги, полилоги) представлены также в изданиях «Живая речь уральского города: Тексты» (1995), «Город в зеркале своего языка» (1996), «Воспоминания работницы М. Н. Колтаковой “Как я прожила жизнь”» (1997), «Разговорная речь носителей массовой городской культуры (на материале г. Омска)» (2007) и многих других.

Разнообразные материалы устной монологической речи представлены в проекте А. А. Кибрика, В. И. Подлесской и др. «Рассказы о сновидениях и другие корпуса звучащей речи».

В целом подавляющее большинство коллекций русских устных спонтанных монологов собрано в рамках единого коммуникативного сценария, чаще всего — «нарративы на заданную тему».

Если обратиться к западному опыту исследования устных спонтанных монологов, то он также весьма разнообразен и опирается на многочисленные корпусы и лингвистические базы данных. В основном это материалы интервью, в том числе записанные в лабораторных условиях, например BACKBONE (записи интервью с нативными носителями английского, французского, немецкого, польского, испанского, турецкого и ненативными носителями английского языка), SLX Corpus of Classic Sociolinguistic Interviews (на материале английского языка), TAUS (фиксирует речь жителей Осло, собранную в 1970-х гг.), ELISA (включает интервью носителей английского языка как второго) и др. По разнообразию методик получения речевого материала можно выделить Lincoln Laboratory Speech Enhancement Corpus (LLSEC), собранный в рамках целого ряда сценариев, моделирующих широкую вариативность условий речи. Крупные корпусы устной речи зачастую включают материалы как устных диалогов, так и монологов: это и Lancaster/IBM Spoken English Corpus (SEC), и The Wellington Corpus of Spoken New Zealand English, и ANDOSL (Australian National Database of Spoken Language), и ELFA (English as a Lingua Franca in Academic Settings), и MICASE (Michigan Corpus of Academic Spoken English).

В том числе можно особо выделить направление исследования устных академических монологов — в основном на материале английского языка (см. работы S. E. Thompson о структуре текста, интонации академических лекций, работы J. Rendle-Short о паузах в академическом монологе, работы M. Cribb о дискурсивных маркерах, интонации и других явлениях в текстах ненативных носителей английского языка и др.), направление исследования детских монологов (работы L. J. Harriet).

Важность любого научного ресурса для исследователей определяется не в последней степени его доступностью для пользователей. К сожалению, онлайн-ресурсов, представляющих материалы русской устной монологической речи, в настоящее время несоизмеримо меньше, чем посвященных этой теме научных работ.

Параллельно со сбором речевого материала коллекции CAT проводится анализ монологической русской речи. В результате был получен целый ряд наблюдений, связанных со спецификой устной спонтанной, в первую очередь монологической, русской речи.

В *общетеоретическом аспекте* обсуждалось, в частности, само понятие спонтанности речи и было предложено два подхода к его трактовке. Прежде всего, спонтанность может пониматься фактически как синоним неподготовленности, и спонтанной в этом случае признается речь неподготовленная, непринужденная и осуществляемая в неофициальной обстановке. Бóльшая часть

проведенных исследований основана именно на таком понимании спонтанности и спонтанной речи. В другой трактовке спонтанность отражает не только и не столько предварительную неподготовленность речи, сколько плохую ее согласованность с конкретными условиями речевой коммуникации, которые порождаются и определяются не в последнюю очередь личностью говорящего. Спонтанность в таком понимании является маркером не первичности, а несогласованности мысли и речи с условиями коммуникации. В материалах собранной коллекции монологических текстов можно найти примеры проявления спонтанности и такого рода.

Столь же общетеоретическими можно считать и поиск единиц описания устной речи, анализ способов сокращения и приращения устного текста, установление градаций (степеней) естественности спонтанной речи и некоторые другие вопросы, включая принципы формирования речевой коллекции и разработку способов балансировки материала как в лингвистическом, так и в психо- и социолингвистическом отношении. Поворот от исследования по преимуществу лабораторной устной речи к более естественным ее типам, порождаемым в условиях повседневной коммуникации, с учетом всех возможных корреляций лингвистических параметров материала с характеристиками говорящего и типом спонтанного монолога можно считать отличительной чертой предлагаемого подхода к анализу устной речи.

На *фонетическом уровне* анализировались паузы хезитации разного типа, являющиеся неотъемлемым свойством спонтанного речепорождения и возникающие во всех видах речи и у всех говорящих; фонетические ошибки, в первую очередь в чтении; а также темп речи — разных говорящих и в разных типах речи.

Лексический аспект исследования представлен, в частности, анализом лексических трансформаций исходных текстов при пересказе и детальной проработкой всех типов эзоединиц, противопоставленных эндоединицам репродуктивов; а также анализом лексических ошибок при чтении. Кроме того, лексический аспект исследования представлен анализом новых, специфичных явлений, свойственных именно повседневной спонтанной речи: новых слов, новых значений или новых коннотаций старых слов, особенностей сочетаемости тех или иных лексических единиц или даже новых идиом, отличающих нашу спонтанную речь; наконец, анализом проявлений внутриязыковой интерференции — в первую очередь между профессиональной и бытовой речью говорящего индивида. Такие проявления, как показали исследования, обнаруживаются прежде всего именно на лексическом уровне.

Морфологический уровень представлен анализом грамматических ошибок в чтении, номинативной лексики в монологах-описаниях, функционирования в разных видах текстов и в речи разных групп информантов различных глагольных форм (инфинитива, причастия и деепричастия), а также некоторыми другими частными наблюдениями морфологического характера.

На *синтаксическом уровне* были описаны синтаксические трансформации исходных текстов при пересказе, структура предикативных единиц, длина текста

в «предложениях» и длина «предложения» в словах (также по преимуществу в репродуктиве, в сравнении с исходными текстами-стимулами) (о способе вычленения в устном тексте единиц, соответствующих предложению, см.: [Bogdanova-Beglarian 2017]). Кроме того, были описаны вставные конструкции и способы передачи чужой речи в разных типах текстов, дана общая синтаксическая характеристика речи говорящих с разным уровнем речевой компетенции, описано функционирование в спонтанных монологах изолированного номинатива.

На *дискурсивном уровне* проанализированы повторы, перебивы, случаи самокоррекции, коммуникативные установки и стратегии говорящего в разных типах текстов, элементы метакоммуникации, отличающие именно спонтанную речь и зависящие как от типа текста, так и от характеристик говорящего. Особенно подробно описаны сценарный и композиционный уровни построения устного монолога-описания и специальные конструкции описательного дискурса.

Лингвистический характер полученных наблюдений (см. [Звуковой корпус… 2013, 2014, 2015]) сопряжен с психо- и социолингвистическим аспектом описания: почти во всех случаях сделана попытка установления корреляции между лингвистическими и экстралингвистическими параметрами материала. В паралингвистическом аспекте были рассмотрены функции смеха и вздохов в спонтанной монологической речи. Прагматический аспект исследования представлен серией методических разработок с использованием материалов Звукового корпуса и предназначенных для преподавания русского языка как иностранного; а также попыткой описания идиолекта — речи конкретной языковой личности, записанной многократно во всех типах коммуникативных ситуаций.

Уже из этого перечня видно, что научным коллективом исследователей — разработчиков CAT — получены результаты действительно многоаспектного описания русской спонтанной речи, в том числе по четырем основным коммуникативным сценариям нашей повседневной жизни — чтению, пересказу, описанию изображения и свободному нарративу. Однако большинство результатов анализа спонтанных монологов было получено лишь на материале изолированных компонентов коллекции CAT, системный анализ и описание монологической речи на всем корпусе данных никогда ранее не проводился — препятствием этому была техническая разнородность элементов коллекции. Поэтому перспективной задачей коллектива ставится не только унификация форматов и стандартов коллекции CAT, но и формирование на ее базе уникальной информационно-исследовательской системы анализа монологической речи, на основе которой можно будет не только проверить на представительном речевом материале выдвинутые ранее научные гипотезы, но и проводить ряд новых исследований, в том числе квантитативных.

В заключение скажем, что результаты таких исследований важны как для собственно лингвистики, так и для ряда смежных научных дисциплин: акустики речи, прагматики, семантики, психолингвистики, социолингвистики, когнитивной лингвистики, исследований дискурса, антропологии, культурологии, этнолингвистики.

Литература

Богданова Н. В. О корпусе текстов живой речи: новые поступления и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии. Вып. 9 (16). По материалам международной конференции «Диалог» (2010) / гл. ред. А. Е. Кибрик. М.: РГГУ, 2010. С. 35–40.

Богданова Н В. Метакоммуникация в устной спонтанной речи (диалог *vs.* монолог) // Коммуникация в социально-гуманитарном знании, экономике, образовании. Материалы III Международной научно-практической конференции. 29–31 марта 2012 г., Минск. Научное электронное издание. Минск, 2012. С. 330–331.

Богданова Н. В., Бродт И. С., Куканова В. В., Павлова О. В., Сапунова Е. М., Филиппова Н. С. О «корпусе» текстов живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / гл. ред. А. Е. Кибрик. М.: РГГУ, 2008. С. 57–61.

Богданова-Бегларян Н. В., Шерстинова Т. Ю., Зайдес К. Д. Корпус «Сбалансированная Аннотированная Текстотека»: методика многоуровневого анализа русской монологической речи //Анализ разговорной русской речи (АР^3-2017): Труды седьмого междисциплинарного семинара / науч. ред. Д. А. Кочаров, П. А. Скрелин. СПб.: Политехника-принт, 2017а. С. 8–13.

Богданова-Бегларян Н. В., Шерстинова Т. Ю., Блинова О. В., Мартыненко Г. Я. Корпус «Один речевой день» в исследованиях социолингвистической вариативности русской разговорной речи // Анализ разговорной русской речи (АР^3-2017): Труды седьмого междисциплинарного семинара / науч. ред. Д. А. Кочаров, П. А. Скрелин. СПб.: Политехника-принт, 2017б. С. 14–20.

Бродт И. С. Спонтанный монолог в лингвистическом и социолингвистическом аспектах (на материале текстов разного типа): дис. … канд. филол. наук. СПб., 2007. 289 с. (машинопись).

Завадская Ю. О. Оговорки в русской устной спонтанной речи: монолог vs. диалог. Курсовая работа. СПб., 2018. 78 с.

Зайдес К. Д. Метакоммуникативные вставки в русской устной спонтанной речи на родном и неродном языке // Коммуникативные исследования. 2016. № 3 (9). С. 19–35.

Зайдес К. Д. Типология метакоммуникативных единиц русской спонтанной монологической речи // Компьютерная лингвистика и вычислительные онтологии. Вып. 1 (Труды XX Международной объединенной научной конференции «Интернет и современное общество», IMS-2017, Санкт-Петербург, 21–23 июня 2017 г. Сб. научных статей). СПб. : Университет ИТМО, 2017. С. 46–56.

Замковец К. С. Поисковые хезитативы в русской спонтанной речи носителей нидерландского языка // Материалы международного молодежного научного форума «Ломоносов-2018» / отв. ред. И. А. Алешковский, А. В. Андриянов, Е. А. Антипов. Москва, МГУ им. М. В. Ломоносова, 9–13 апреля 2018 г. М.: МГУ, 2018 [Электронный ресурс].

Звуковой корпус как материал для анализа русской речи: коллективная монография. Ч. 1: Чтение. Пересказ. Описание / отв. ред. Н. В. Богданова-Бегларян. СПб.: Филологический ф-т СПбГУ, 2013. 532 с.

Звуковой корпус как материал для анализа русской речи:коллективная монография. Ч. 2: Теоретические и практические аспекты анализа. Т. 1: О некоторых особенностях устной спонтанной речи разного типа. Звуковой корпус как материал для преподавания русского языка в иностранной аудитории / отв. ред. Н. В. Богданова-Бегларян. СПб.: Филологический ф-т СПбГУ, 2014. 396 с.

Звуковой корпус как материал для анализа русской речи: коллективная монография. Ч. 2: Теоретические и практические аспекты анализа. Т. 2: Звуковой корпус как материал для новых лексикографических проектов / отв. ред. Н. В. Богданова-Бегларян. СПб.: Филологический ф-т СПбГУ, 2015. 364 с.

Иванова О. А. Специфика бытовой речи различных профессионально ориентированных групп: дис. … маг. лингв. СПб., 2011. 112 с. (машинопись).

Казак М. В. Паузы хезитации в спонтанной речи на родном и неродном языках (на материале речи франкофонов). Saarbrücken: Lambert Academic Publishing, 2015. 74 с.

Куканова В. В. Лингвистический анализ репродуцированных текстов (на материале звукового корпуса русской речи юристов): дис. … канд. филол. наук. СПб., 2009. 328 с. (машинопись).

Малиновская А. И. Репродуктив как объект многоаспектного анализа (на материале Звукового корпуса русского языка): дис. … маг. лингв. СПб., 2012. 103 с. (машинопись).

Метлова В. А. Темп речи и паузы хезитации в речи на родном и неродном языках: монография. Saarbrücken: Palmarium Academic Publishing, 2013. 79 с.

Русская спонтанная речь. Свободные монологи-рассказы на заданную тему. Тексты. Лексические материалы / сост. В. В. Куканова; отв. ред. и авт. предисл. Н. В. Богданова. СПб.: Филологический ф-т СПбГУ, 2008. 208 с.

Русская спонтанная речь. Монологи-репродуктивы. Тексты. Лексические материалы / сост. В. В. Куканова; отв. ред. и авт. предисл. Н. В. Богданова. СПб.: Филологический ф-т СПбГУ, 2010. 132 с.

Русская спонтанная речь. Монологи-описания. Тексты. Лексические материалы / сост. В. В. Куканова; отв. ред. и авт. предисл. Н. В. Богданова. СПб.: Филологический ф-т СПбГУ, 2011. 140 с.

Русская спонтанная речь. Спонтанные монологи разных типов. Тексты. Лексические материалы (CD) / сост. Н. В. Богданова-Бегларян, И. С. Бродт; отв. ред. М. Краузе // Бюллетень Фонетического Фонда. Бохум (Германия), 2020. 100 с. (в печати).

Русский язык повседневного общения: особенности функционирования в разных социальных группах: коллективная монография / отв. ред. Н. В. Богданова-Бегларян. СПб.: ЛАЙКА, 2016. 244 с.

Сапунова Е. М. Неподготовленное чтение как вид речевой деятельности и тип устного спонтанного монолога (на материале русского языка): дис. … канд. филол. наук. СПб., 2009. 237 с. (машинопись).

Степихов А. А. Соотношение синтаксического и интонационного членения в спонтанном монологе: дис. … канд. филол. наук. СПб., 2005. 197 с. (машинопись).

Филиппова Н. С. Принципы построения устного описательного дискурса (на материале русской спонтанной речи): дис. … канд. филол. наук. СПб., 2010. 220 с. (машинопись).

Хан Н. А. Устные спонтанные монологи разного типа в коммуникативно-дискурсивном аспекте (на материале Звукового корпуса русского языка): дис. … канд. филол. наук. СПб., 2013. 277 с. (машинопись).

Чэн Чэнь. Хезитации в русской устной речи носителей китайского языка: дис. … канд. филол. наук. СПб., 2018. 205 с. (машинопись).

Asinovsky, A., Bogdanova, N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation // Text, Speech and Dialogue. 12th International Conference, TSD 2009. Proceedings / eds. V. Matoušek, P. Mautner.Pilsen, Czech Republic, September 2009. P. 250–257.

Bogdanova-Beglarian N. V. In Search of Phrase Boundaries in Spontaneous Speech // *SPECOM 2017.* Lecture Notes in Artificial Intelligence, LNAI. Vol. 10458. Springer, Switzerland, 2017. P. 456–463.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // Speech and Computer. 18th International Conference, SPECOM 2016. Budapest, Hungary, August, 23–27, 2016a. Proceedings / eds. A. Ronzhin, R. Potapova, G. Németh. P. 659–666.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G. An Exploratory Study on Sociolinguistic Variation of Spoken Russian // Speech and Computer. 18th International Conference, SPECOM 2016. Budapest, Hungary, August, 23–27, 2016b. Proceedings / eds. A. Ronzhin, R. Potapova, G. Németh. P. 100–107.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G. Linguistic Features and Sociolinguistic Variability in Everyday Spoken Russian // *SPECOM 2017.* Lecture Notes in Artificial Intelligence, LNAI. Vol. 10458. Springer, Switzerland, 2017. P. 503–511.

N. V. Bogdanova-Beglarian, O. V. Blinova, K. D. Zaides, T.Ju. Sherstinova
Saint-Petersburg State University (Russia, Saint-Petersburg)
n.bogdanova@spbu.ru,o.blinova@spbu.ru, kristina.zaides@student.spbu.ru,
t.sherstinova@spbu.ru

CORPUS "BALANCED ANNOTATED TEXT COLLECTION (TEXTOTEC)" (SAT): STUDYING THE SPECIFICITY OF RUSSIAN MONOLOGICAL SPEECH

The article represents one of the Russian speech corpora: a collection of monologic texts, known as the "Balanced Annotated Text Collection (Textotec)" (SAT). This corpus

was being assembled in St. Petersburg State University for more than 20 years, using the author's (N. V. Bogdanova-Beglarian's) methodology of data collection, which involves a fairly strict set of experimental procedures. SAT is designed to study various types of spontaneous monologues(reading, retelling, image description, story on the topic) and it contains texts recorded from five professionally-oriented groups of native speakers (medical doctors, lawyers, computer specialists, philologists, teachers of Russian as a foreign language, and teachers-philosophers), several blocks of students speech (philologists and non-philologists), as well as four blocks of the interfered Russian speech of native speakers of other languages: Americans, Chinese, Francophone and Dutch. In total, there are about 700 texts in the SAT and about 50 hours of sound recording. In the article, against the background of other Russian-speaking and foreign speaking corpora, a description of this linguistic resource is given, the main topics developed on its material are marked, and prospects for continuing work are outlined.

Keywords: modern Russian language, oral monologic speech, speech corpus, natural language processing, database, linguistic experiment, reading, description of the image, retelling of the text (reproductive), spontaneous monologue, sociolinguistics, psycholinguistics.

References

Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation. *Text, Speech and Dialogue. 12th International Conference, TSD 2009. Proceedings*. Eds. V. Matoušek, P. Mautner. Pilsen, Czech Republic, September 2009,pp. 250–257.

Bogdanova N. V. [Metacommunication in Oral Spontaneous Speech (Dialogue vs. Monologue)]. *Kommunikatsiya v sotsial'no-gumanitarnomznanii, ekonomike, obrazovanii. Materialy III Mezhdunarodnoi nauchno-prakticheskoi konferentsii.* 29–31 marta 2012 g. [Communication in Socio-Humanitarian Knowledge, Economics, Education. Materials of the III International Scientific and Practical Conference. 29–31 March, 2012]. Minsk, Scientific electronic publication, 2012, pp. 330–331. (In Russ.)

Bogdanova N. V. [On the Corpus of Texts of Living Speech: New Acquisitions and First Results of the Research]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam mezhdunarodnoi konferentsii "Dialog" (2010)* [Computer Linguistics and Intellectual Technologies. Based on the Materials of the International Conference "Dialogue" (2010)]. Iss. 9 (16). Moscow, 2010, pp. 35–40. (In Russ.)

Bogdanova N. V., Brodt I. S., Kukanova V. V., Pavlova O. V., Sapunova E. M., Filippova N. S. [On the Corpus of Texts of Living Speech: the Principles of Formation and the Possibility of Description]. *Komp'yuternaya lingvistika I intellektual'nye tekhnologii. Po materialam ezhegodnoi mezhdunarodnoi konferentsii "Dialog" (2008)* [Computer Linguistics and Intellectual Technologies. Based on the Materials of the Annual International Conference "Dialogue" (2008)]. Iss. 7 (14). Moscow, 2008, pp. 57–61. (In Russ.)

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Baeva E., Martynenko G., Ryko A. [Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech]. *Speech and Computer. 18th International Conference, SPECOM 2016*. Budapest, Hungary, August, 23–27, 2016a. Proceedings. Eds. A. Ronzhin, R. Potapova, G. Németh, pp. 659–666.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G. [An Exploratory Study on Sociolinguistic Variation of Spoken Russian]. *Speech and Computer. 18th International Conference, SPECOM 2016*. Budapest, Hungary, August, 23–27, 2016b. Proceedings. Eds. A. Ronzhin, R. Potapova, G. Németh, pp. 100–107.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G. [Linguistic Features and Sociolinguistic Variability in Everyday Spoken Russian]. *SPECOM 2017. Lecture Notes in Artificial Intelligence, LNAI*, vol. 10458. Springer, Switzerland, 2017, pp. 503–511.

Bogdanova-Beglarian N. V. (ed.) [*Everyday Russian Language in Different Social Groups.* Collective. Monograph]. *Russkii yazyk povsednevnogo obshcheniya: osobennosti funktsionirovaniya v raznykh sotsial'nykh gruppakh. Kollektivnaya monografiya*. St. Petersburg, 2016. 244 p. (In Russ.).

Bogdanova-Beglarian N. V. (ed.) *Zvukovoi korpus kak material dlya analiza russkoi rechi. Kollektivnaya monografiya. Ch. 1. Chtenie. Pereskaz. Opisanie* [Speech Corpus as a Base for Analysis of Russian Speech. Collective Monograph. Part 1. Reading. Retelling. Description]. St. Petersburg, 2013. 532 p. (In Russ.)

Bogdanova-Beglarian N. V. (ed.) *Zvukovoi korpus kak material dlya analiza russkoi rechi. Kollektivnaya monografiya. Ch. 2. Teoreticheskie i prakticheskie aspekty analiza. T. 1. O nekotorykh osobennostyakh ustnoi spontannoi rechi raznogo tipa. Zvukovoi korpus kak material dlya prepodavaniya russkogo yazyka v inostrannoi auditoria* [Speech-Corpus as a Base for Analysis of Russian Speech. Collective Monograph. Part 2. Theoretical and Practical Aspects of Analysis. Vol. 1. On Some Features of Oral Spontaneous Speech of Different Types. Speech Corpus as a Material for Teaching Russian in a Foreign Audience]. St. Petersburg, 2014. 396 p. (In Russ.)

Bogdanova-Beglarian N. V. (ed.) *Zvukovoi korpus kak material dlya analiza russkoi rechi. Kollektivnaya monografiya. Ch. 2. Teoreticheskie i prakticheskie aspekty analiza. T. 2. Zvukovoi korpus kak material dlya novykh leksikograficheskikh proektov* [Speech-Corpus as a Base for Analysis of Russian Speech. Collective Monograph. Part 2. Theoretical and Practical Aspects of Analysis. Vol. 2. Speech Corpus as a Material for New Lexicographic Projects]. St. Petersburg, 2015. 364 p. (In Russ.)

Bogdanova-Beglarian N. V. [In Search of Phrase Boundaries in Spontaneous Speech]. *SPECOM 2017. Lecture Notes in Artificial Intelligence, LNAI*, vol. 10458. Springer, Switzerland, 2017, pp. 456–463.

Bogdanova-Beglarian N. V., Brodt I. S. (orig.), Krause M. (ed.) [Russian Spontaneous Speech. Spontaneous Monologues of Different Types. Texts. Lexical Materials (CD)]. *Byulleten'Foneticheskogo Fonda* [Bulletin of the Phonetic Fund]. Bochum (Germany), 2020. 100 p. (In Russ.) (In Print.)

Bogdanova-Beglarian N. V., Sherstinova T.Ju., Blinova O. V., Martynenko G.Ja. [Corpus "One Speaker's Day" in Studies of Sociolinguistic Variability of Russian Colloquial

Speech]. *Analiz razgovornoi russkoirechi (AR³-2017): Trudy sed'mogo mezhdistsiplinarnogo seminara* [Analysis of Spoken Russian Speech (AR³-2017): Proceedings of the 7th Interdisciplinary Seminar]. St. Petersburg, 2017b, pp. 14–20. (In Russ.)

Bogdanova-Beglarian N. V., Sherstinova T.Ju., Zaides K. D. [Corpus "Balanced Annotated Text Library": Methodology Multi-level Analysis of the Russian Monologue Speech]. *Analiz razgovornoi russkoi rechi (AR³-2017): Trudy sed'mogo mezhdistsiplinarnogo seminara* [Analysis of Spoken Russian Speech (AR³-2017): Proceedings of the 7th Interdisciplinary Seminar]. St. Petersburg, 2017a, pp. 8–13. (In Russ.)

Brodt I. S. *Spontannyi monolog v lingvisticheskom i sotsiolingvisticheskom aspektakh (na material tekstov raznogo tipa).* Dis. kand. filol. nauk [Spontaneous Monologue in Linguistic and Sociolinguistic Aspects (Based on the Material of Texts of Different Types). Dis. cand. philol. sci.]. St. Petersburg, 2007. 289 p. (In Russ.)

Cheng Chen. *Khezitatsii v russkoi ustnoi rechi nositelei kitaiskogo yazyka.* Dis. kand. filol. nauk [Hesitations in the Russian spoken language of Chinese speakers. Dis. cand. philol. sci.]. St. Petersburg, 2018. 205p. (In Russ.)

Filippova N. S. *Printsipy postroeniya ustnogo opisatel'nogo diskursa (na materiale russkoi spontannoi rechi).* Dis. kand. filol. nauk [Principles of Constructing of Oral Descriptive Discourse (Based on Russian Spontaneous Speech). Dis. cand. philol. sci.]. St. Petersburg, 2010. 220 p.(In Russ.)

Ivanova O. A. *Spetsifika bytovoi rechi razlichnykh professional'no orientirovannykh grupp.* Dis. mag. lingv. [Specificity of Everyday Speech of Various Professionally Oriented Groups. Dis. mast. ling.]. St. Petersburg, 2011. 112 p. (In Russ.)

Kazak M. V. *Pauzy khezitatsii v spontannoi rechi na rodnom i nerodnom yazykakh (na material rechi frankofonov)* [Pauses of Hezitations in Spontaneous Speech in Native and Non-Native Languages (on the Material of the Speech of Francophone)]. Saarbrücken, Lambert Academic Publishing, 2015. 74 p. (In Russ.)

Khan N. A. *Ustnye spontannye monologi raznogo tipa v kommunikativno-diskursivnom aspekte (na material Zvukovogo korpusa russkogo yazyka).* Dis. kand. filol. Nauk [Spoken Spontaneous Monologues of Various Types in Communicative and Discourse Aspects (Based on the Spoken Corpus of Russian Language). Dis. cand. philol. sci.]. St. Petersburg, 2013. 277p.

Kukanova V. V. (orig.), Bogdanova N. V. (ed.) *Russkaya spontannaya rech'. Monologi-opisaniya. Teksty. Leksicheskie materialy* [Russian Spontaneous Speech. Description Monologues. Texts. Lexical Materials]. St. Petersburg, 2011. 140 p. (In Russ.)

Kukanova V. V. (orig.), Bogdanova N. V. (ed.) *Russkaya spontannaya rech'. Svobodnye monologi-rasskazy na zadannuyu temu. Teksty. Leksicheskie materialy* [Russian Spontaneous Speech. Free Monologues-Stories on a Topic. Texts. Lexical Materials]. St. Petersburg, 2008. 208 p. (In Russ.)

Kukanova V. V.(orig.), Bogdanova N. V. (ed.) *Russkaya spontannaya rech'. Monologi-reproduktivy. Teksty. Leksicheskie materialy* [Russian Spontaneous Speech. Monologues-Reproductives. Texts. Lexical Materials]. St. Petersburg, 2010. 132 p. (In Russ.)

Kukanova V. V. *Lingvisticheskii analiz reprodutsirovannykh tekstov (na material zvukovogo korpusa russkoi rechi yuristov).* Dis. kand. filol. Nauk [Linguistic Analysis of

Reproduced Texts (Based on the Material of the Russian speech corpus of Lawyers). Dis. cand. philol. sci.]. St. Petersburg, 2009. 328 p. (In Russ.)

Malinovskaya A. I. *Reproduktiv kak ob'ekt mnogoaspektnogo analiza (na materiale Zvukovogo korpusa russkogo yazyka).* Dis. mag. lingv. [Reproductive as an Object of Multi-Aspect Analysis (Based on the Material of the Speech Corpus of Russian). Dis. mast. ling.]. St. Petersburg, 2012. 103 p. (In Russ.)

Metlova V. A. *Temp rechi i pauzy khezitatsii v rechi na rodnom I nerodnom yazykakh. Monografiya* [The Rate of Speech and Pause of the Hezation in Speech in Native and Non-Native Languages. Monograph]. Saarbrücken, Palmarium Academic Publishing, 2013. 79 p. (In Russ.)

Sapunova E. M. *Nepodgotovlennoe chtenie kak vid rechevoi deyatel'nosti i tip ustnogo spontannogo monologa (na materiale russkogo yazyka).* Dis. kand. filol. nauk [Unprepared Reading as a Type of Speech Activity and Type of Oral Spontaneous Monologue (Based on the Material of Russian Language). Dis. cand. philol. sci.]. St. Petersburg, 2009. 237 p. (In Russ.)

Stepikhov A. A. *Sootnoshenie sintaksicheskogo i intonatsionnogo chleneniya v spontannom monologe.* Dis. kand. filol. Nauk [The Correlation of Syntactic and Intonation Partitioning in Spontaneous Monologues. Dis. cand. philol. sci.]. St. Petersburg, 2005. 197 p. (In Russ.)

Zaides K. D. [Metacommunication Inserts in Russian Oral Spontaneous Speech in Native and Non-Native Language]. *Kommunikativnye issledovaniya* [Communicative Research], 2016, no. 3 (9), pp. 19–35. (In Russ.)

Zaides K. D. [Typology of Metacommunicative Units of Russian Spontaneous Monologic Speech]. *Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vyp. 1 (Trudy XX Mezhdunarodnoi ob'edinennoi nauchnoi konferentsii "Internet i sovremennoe obshchestvo", IMS-2017, Sankt-Peterburg, 21–23 iyunya 2017 g. Sb. nauchnykhstatei)* [Computer Linguistics and Computing Ontologies. Iss. 1 (Proceedings of the XX International Joint Scientific Conference "Internet and Contemporary Society", IMS-2017, St. Petersburg, June 21–23, 2017, Collection of Scientific Articles)]. St. Petersburg, 2018, pp. 46–56. (In Russ.)

Zamkovec K. S. [Search Hesitatives in the Russian Spontaneous Speech of Native Speakers of the Netherlands]. *Materialy mezhdunarodnogo molodezhnogo nauchnogo foruma "Lomonosov-2018". Moskva, MGU im. M. V. Lomonosova, 9–13 aprelya 2018 g.* [Proceedings of the International Youth Scientific Forum "Lomonosov-2018". Moscow, Moscow State University M. V. Lomonosov. April 9–13, 2018]. Moscow, 2018 [Electronic resource].(In Russ.)

Zavadskaya Yu.O. *Ogovorki v russkoi ustnoi spontannoi rechi: monolog vs. dialog. Kursovaya rabota* [Bloopers in Russian Oral Spontaneous Speech: Monologue vs. Dialogue. Student course work]. St. Petersburg, 2018. 78 p. (In Russ.)