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

КОРПУС РУССКОГО ЯЗЫКА ПОВСЕДНЕВНОГО ОБЩЕНИЯ «ОДИН РЕЧЕВОЙ ДЕНЬ» (ОРД): ТЕКУЩЕЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ*

В статье описываются этапы создания и расширения корпуса повседневной русской речи «Один речевой день» (ОРД), а также требования к публикации его онлайн-версии. Методологическая основа ОРД — осуществление звукозаписей в максимально естественных условиях. Для участия отбирались информанты-добровольцы, готовые прожить день «с диктофоном на шее» и заполнить ряд анкет. Перед аудиозаписью информанты проходили инструктаж, на котором учились пользоваться звукозаписывающей техникой и получали «Памятку информанта». Мы просили вмешиваться в процесс звукозаписи как можно меньше и вести себя «как обычно». Все записи осуществлялись анонимно, информанты и их собеседники обозначались кодами, а частная информация в транскрипте специальным образом маркировалась. В результате материалы корпуса содержат преимущественно разговоры из частной жизни информантов. Что же касается требований к публикации материалов онлайн, важным является сохранение анонимности авторства речевого материала. Таким образом, транскрипты звукозаписей могут быть опубликованы только при условии анонимизации личных имен, фамилий, прозвищ, а также исключения из текстов, представленных на сайте, любой другой информации, которая может повлечь раскрытие личности говорящего. В статье предлагается способ анонимизации личных данных и ставится проблема цензурной редактуры транскриптов.

Ключевые слова: русский язык, устная повседневная речь, речевой корпус, многоуровневое аннотирование, личная информация, подготовка корпуса к публикации.

* При поддержке гранта РНФ (проект №18-18-00242 «Система прагматических маркеров русской повседневной речи»).

1. Введение

Корпус повседневной русской речи «Один речевой день» (ОРД) создается, пополняется и активно анализируется на филологическом факультете СПбГУ с 2007 г.[1] Целью создания корпуса является изучение устной речи, бытовой и профессиональной коммуникации. Методологическая основа ОРД — осуществление звукозаписей в условиях, максимально приближенных к естественным, для чего используется методика многочасового мониторинга, непрерывной 24-часовой записи всей речевой продукции информантов (см. [Asinovsky et al. 2009; Богданова-Бегларян и др. 2017]). Подобная методика сбора речевого материала используется в японских лингвистических исследованиях (см. об этом, например: [Сибата 1983; Campbell 2004]); кроме того, она применялась при создании демографического подкорпуса Британского национального корпуса [Burnard 2007].

Сегодня корпус ОРД содержит около 1250 часов звучания, более 2800 коммуникативных макроэпизодов, включает материалы речи 128 информантов (из них 68 мужчин и 60 женщин в возрасте от 18 до 83 лет), а также более 1000 их основных коммуникантов. Объем текстовых расшифровок корпуса превышает 1 млн словоупотреблений (см. о нем подробнее: [Русский язык… 2016]). В настоящей статье описываются этапы создания и расширения корпуса ОРД и его текущее состояние, а также возможные условия публикации его онлайн-версии.

2. Технология создания корпуса

2.1. Отбор и инструктаж информантов

Для участия в записи отбирались информанты-добровольцы, готовые прожить день «с диктофоном на шее» и заполнить ряд анкет. При отборе учитывался возрастной фактор (принять участие в записи могли лишь лица, достигшие 18 лет) и фактор родного языка (русский язык для информанта должен быть родным и основным). К участию в эксперименте приглашались исключительно информанты, живущие в городской среде.

Перед аудиозаписью добровольцы проходили инструктаж, в ходе которого учились пользоваться звукозаписывающей техникой, и получали «Памятку информанта», где отражены основные требования к организационным и техническим особенностям проведения исследования. В «Памятке» указывалось, в частности, что диктофон следует отключать только на время смены батареек (приблизительно раз в 6 часов). Мы также просили информантов по возможности выключать источники фоновых шумов (телевизор, радио и др.) и обязательно предупреждать собеседников о проводимой звукозаписи. Перед началом записи информанты подписывали «Согласие», подготовленное юридической службой СПбГУ.

[1] Работа по созданию корпуса была начата при поддержке гранта РГНФ (проект №07-04-94515е/Я).

2.2. Методика звукозаписи

Первые фонограммы корпуса ОРД собраны с помощью цифрового диктофона WS-320M. В ходе расширения корпуса, выполненного в рамках проекта РНФ 2014–2016 гг.[2], запись производилась с помощью профессиональных диктофонов Roland R09-HR с внешними конденсаторными микрофонами SONY ECM-T140 в формате PCM (WAV) 44 100Гц, 16 бит, стерео.

Мы стремились, чтобы в корпус ОРД вошли «обычные дни» наших информантов (без редких событий или из ряда вон выходящих происшествий). Мы просили также вмешиваться в процесс звукозаписи как можно меньше и вести себя как обычно. В упомянутой выше «Памятке информанта» оговорены основные режимы звукозаписи — «стационарный» и «мобильный». В «стационарном» режиме — при длительном нахождении в одном помещении — мы рекомендовали снимать диктофон и размещать его вблизи себя на какой-то поверхности (например, на столе), а выносной микрофон отключать. В «мобильном» режиме, предполагающем активные перемещения информанта, мы советовали вести запись с помощью петличного микрофона.

2.3. Анкетирование информантов, социологическая информация в базе данных

В ходе записи все информанты вели «Дневник речевого дня». Это шаблон, в котором фиксируются основные события, произошедшие в течение дня, и описываются основные коммуниканты. «Дневник» содержит четыре поля: «время» (когда происходил разговор), «место» (где происходил разговор), «собеседники» (с кем происходил разговор), «вид деятельности» (что сопровождало разговор).

Кроме того, в ходе записи проводилось анкетирование: информанты корпуса заполняли одну социологическую анкету и проходили три психологических теста.

Социологическая анкета, в которую записывались данные информантов и его основных собеседников (коммуникантов), разработана на основе данных Федеральной службы городской статистики и традиционных для социолингвистики параметров описания говорящих [Bogdanova-Beglarian et al. 2016]. В анкете отражены пол, возраст, место рождения и наиболее длительного проживания информантов и др. Информация о каждом коммуниканте дополнена его социальной ролью по отношению к информанту.

Все данные поступают в информационную базу данных формата MS Access. Таблицы с данными об информантах и коммуникантах, взятыми из социологических анкет, содержат следующие основные поля: (1) код информанта, (2) пол информанта, (3) возраст информанта на момент записи, (4) место рождения, (5) родной язык, (6) другие языки, которыми владеет информант, (7) национальность

[2] Грант РНФ № 14-18-02070 «Русский язык повседневного общения: особенности функционирования в разных социальных группах».

родителей (заполняется по желанию), (8) социальное происхождение, (9) уровень образования (среднее специальное, высшее и т. п.), (10) квалификация (специальность) по диплому, (11) прошлые профессии или опыт работы, (12) профессия или род деятельности на момент записи, (13) места наиболее длительного проживания, (14) комментарии [Русский язык… 2016].

Записи велись анонимно, все информанты и их собеседники обозначены кодами. Код в формате S01, S10, S100 присваивается в зависимости от порядкового номера информанта. Коммуникантам (мужчинам, женщинам и детям) присваиваются коды типа М1, М2, Ж1, Ж2, Р1, Р2 — в зависимости от момента их появления в эпизоде коммуникации (звуковом файле и файле расшифровки со стандартным названием типа ordS01-nn, где 01 — код информанта, nn — номер макроэпизода «речевого дня»). Возможны коды НМ или НЖ — в случае, когда идентифицировать коммуниканта по записи не удается.

Три стандартных психологических теста (анкеты), которые выполняли участники эксперимента, — это тест Г. Айзенка, тест FPI и тест Р. Кеттела, с бланками ответов, подготовленными для заполнения. Психологическая информация представлена для 85 информантов корпуса (это звукозаписи, выполненные в 2014–2016 гг.).

3. Обработка речевого материала

3.1. Форматирование файлов звукозаписи. Членение на «макроэпизоды»

Форматирование файлов ОРД подразумевает преобразование первичного формата записи в формат PCM (22 050 Гц, 16 бит, моно), удаление пауз продолжительностью более 2–5 минут и членение исходных файлов звукозаписи на файлы, содержащие отдельные коммуникативные макроэпизоды.

Коммуникативные макроэпизоды — это фрагменты «речевого дня», объединенные местом, условиями и участниками коммуникации. Эпизоды описываются по определенной схеме с учетом параметров «тип коммуникации», «условия коммуникации», «социальные роли говорящих», «место коммуникации» (см. [Шерстинова 2013]).

3.2. Техника расшифровки, система аннотирования

Расшифровка звукозаписей выполняется в программе ELAN [ELAN] и предполагает заполнение восьми базовых уровней: Frase (реплики говорящих), Speaker (код говорящего), Events (невербальные аудиособытия), Voice (качество голоса говорящего), FonetCom (фонетический комментарий), FraseComment (фразовый комментарий), Notes (общий комментарий), Episode (мини-эпизод речевой коммуникации) (подробнее см. [Sherstinova 2015; Русский язык… 2016]).

Уровень «Frase» — основной. Он содержит транскрипт звукозаписи, на нем производится членение звукозаписи на сегменты (боксы), не содержащие речевого

сигнала (паузы, обозначаемые символом <*П>), и реплики говорящих. Транскрипт выполняется в стандартной орфографии.

Некоторые отступления от правил стандартной орфографии допускаются при расстановке пробелов, оформлении дефисных написаний. Эти отступления введены для того, чтобы избежать ряда проблем, влияющих на качество дальнейшей обработки текстов (лемматизации, морфологического анализа и др.) (см. [Блинова 2015]). Рекомендации для расшифровщиков требуют, чтобы цепочки символов, которые считаются отдельными токенами, были отделены пробелами с обеих сторон. Например, таким образом оформляются: словоформы, символы, обозначающие паралингвистические явления, паузы, наложения речи, знаки фразового и синтагматического членения </>, <//> и др. Для неоднословных выражений введено одно правило: названия из двух и более слов объединяются с помощью нижнего подчеркивания (например, *От_заката_до_рассвета*).

Инструкция для расшифровщика рекомендует оформлять написания, в графическом представлении которых задействован дефис, следующим образом: <* *то*, * *таки*> даются как отдельные слова, за исключением неопределенных местоимений и местоименных наречий (*где-то*, *куда-то* и т. д.), а также *всё-таки*; сложные слова с дефисом (за исключением прилагательных, обозначающих цвет, имен собственных и некоторых наречий, имеющих в своей основе дуплеты), а также лексические повторы принимаются за два отдельных слова. В более подробной инструкции оговаривается, что раздельно (через пробел, не по правилам орфографии) пишутся, во-первых, различные дублеты: прилагательные (*быстрый быстрый*), наречия (*долго долго*), глаголы (*сидел сидел*), частицы (*да да*), междометия (*ха ха*) и др.; во-вторых, частицы *-то*, *-ка*, *-таки*, *-де* (*я то знаю, покажи ка, сказал таки*); в-третьих, существительные с приложением (*девушка красавица*); в-четвертых, числительные со значением приблизительного количества (*два три*).

Конвенции дискурсивной транскрипции ОРД подробно описаны в [Шерстинова и др. 2009]. В последнее время список стандартных помет расширен за счет обозначения паралингвистических явлений (<*З> — зевок, <*Ц> — цыканье, <*S> — шмыганье носом, <*G> — гортанные неречевые звуки, <*Г> — причмокивание и др.). Введены новые обозначения для маркирования незавершенных реплик, продолжающихся после паузы. Для обозначения сегментов транскрипта, подлежащих анонимизации (прежде всего личных имен), введена дополнительная помета <%>.

3.3. Подготовка транскриптов к лингвистической разметке

Перед лингвистическим аннотированием (лемматизацией, морфологической разметкой, синтаксической разметкой и др.) файлы расшифровки *.eaf подвергаются ручному экспертному редактированию и автоматической обработке. Во-первых, все файлы обрабатываются с помощью утилиты «Corrector» (собственная разработка научного коллектива), позволяющей автоматически исправить возможные технические огрехи (например, убрать из транскрипта лишние пробелы), а также выявить несоответствия между уровнями «Frase» и «Speaker» в случаях наложения

речи (периодах одновременного говорения) нескольких собеседников. Наложения в повседневной речи наблюдаются достаточно часто и существенно затрудняют ее анализ. В случаях выявления несоответствий аннотаций на указанных уровнях проводится ручная коррекция соответствующих фрагментов расшифровок и их повторный анализ.

Во-вторых, файлы расшифровки *.eaf проходят обработку программой «Eafer» (также собственная разработка коллектива). С ее помощью осуществляется преобразование одноуровневого представления речевого материала, принятого за основу транскрибирования корпуса ОРД с момента его основания в 2007 г., в многоуровневое представление, где каждый речевой уровень относится только к одному говорящему. Этот этап обработки позволяет разделить речевой материал разных говорящих.

В-третьих, выполняется ручная коррекция границ боксов аннотаций для реплик с наложением речи. Коррекция границ аннотаций осуществляется непосредственно в среде ELAN для каждого из уровней, содержащих реплики говорящих. После этого файлы аннотаций формата *.eaf считаются готовыми к автоматической обработке.

4. Подготовка ОРД к публикации: анонимизация данных и цензурная редактура

Звукозаписи корпуса ОРД содержат преимущественно разговоры из частной жизни информантов. Анонимность участников звукозаписи (при открытости их социологических характеристик) — один из ключевых моментов методики сбора данных, который позволял участникам проживать свой «речевой день» свободно и «естественно». Поэтому наиболее важным условием публикации материалов корпуса онлайн является сохранение анонимности авторства речевого материала.

Для исключения атрибуции говорящего по акустическим свойствам голоса и манере речи сами звукозаписи повседневного общения, по-видимому, не могут быть опубликованы в свободном доступе. Что касается транскриптов звукозаписей, то они могут быть опубликованы только при условии полной анонимизации личных имен, фамилий, прозвищ, а также исключения из текстов, представленных на сайте, любой другой информации, которая может повлечь раскрытие личности говорящего (номера телефона, паспортных данных, конкретных мест работы и прочей озвученной информации). Кроме того, не все эпизоды речевой коммуникации могут быть опубликованы по этическим соображениям.

Сегодня актуальными задачами, стоящими перед авторским коллективом, работающим с корпусом ОРД, являются следующие:

1) определение типов эпизодов, которые не могут быть представлены на сайте ни в каком виде, к решению этой задачи будут привлечены квалифицированные юристы;

2) отбор коммуникативных эпизодов, которые могут быть опубликованы после их анонимизации;

3) анонимизация транскриптов: замена всей личной информации — в первую очередь имен и фамилий — на иные, имеющие ту же акцентно-ритмическую структуру;

4) «цензурная» редактура.

Повседневная устная речь содержит заметное количество непечатной лексики. Необходимо принять решение, каким образом такие выражения будут представлены на сайте (как известно, публикация нецензурных слов в сети запрещена). Однако для лингвистических исследований желательно, чтобы по тексту расшифровки можно быть однозначно восстановить произнесенный текст, следовательно, предстоит получить полный список цензурируемых единиц и выработать способы их передачи.

5. Корпус ОРД онлайн

В рамках текущего проекта РНФ коллективом запланирована публикация в сети Интернет со свободным доступом материалов корпуса с возможностью текстового поиска и фильтрации речевого материала по социологическим характеристикам говорящих и условиям коммуникации. Объем подкорпуса, планируемого к публикации, — 300 тыс. словоупотреблений.

Сайт корпуса ОРД будет опубликован в свободном доступе на сервере СПбГУ по адресу http://www.ord-corpus.spbu.ru. Пользовательский интерфейс онлайн-версии корпуса будет обеспечивать текстовый поиск по заданному слову (подстроке). Каждая реплика, полученная по запросу, будет сопровождаться социологической информацией о говорящем (пол, возраст, профессия и др.), а также о типе коммуникативной ситуации (бытовой разговор, профессиональный разговор, учебный разговор и др.). Будет предусмотрена возможность развертывания выдачи до нескольких реплик.

6. Заключение

Корпус ОРД задуман как инструмент мониторинга современной русской речи, направлен на фиксацию материала естественной коммуникации и тем самым — создание условий для научного описания современного повседневного русского языка и анализа особенностей его функционирования в разных социальных группах, в разных ситуациях, в бытовом и профессиональном общении, при паритетных и непаритетных отношениях между говорящими — носителями различных социальных ролей. Корпус достиг значительного для речевого корпуса объема и продолжает развиваться. В настоящее время на его основе разрабатывается система прагматических маркеров повседневной русской речи. Сегодня готовится его публикация на сайте. Мы надеемся, что ОРД станет значимым общедоступным ресурсом для лингвистических исследований самых разных направлений.

Литература

Блинова О. В. «Немашинное обучение»: экспериментальная проверка применимости правил ручной токенизации // Труды международной конференции «Корпусная лингвистика-2015». СПб. : Изд-во СПбГУ, 2015. С. 111–120.

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, 2017. С. 14–20.

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

Сибата Т. Исследования языкового существования в течение 24 часов // Языкознание в Японии / ред. В. М. Алпатов, И. Ф. Вардуль. М. : Радуга, 1983. С. 134–141.

Шерстинова Т. Ю., Степанова С. Б., Рыко А. И. Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Материалы XXXVIII Международной филологической конференции. Секция: «Формальные методы анализа русской речи». СПб. : Изд-во СПбГУ, 2009. С. 66–75.

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 // Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI. Vol. 57292009. Berlin-Heidelberg : Springer, 2009. P. 250–257.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016. Lecture Notes in Artificial Intelligence, LNAI. Vol. 9811. Switzerland : Springer, 2016. P. 659–666.

Burnard L. (ed.) Reference Guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, 2007 [Электронный ресурс]. URL: http://www.natcorp.ox.ac.uk/docs/URG/.

Campbell N. Speech & Expression; the Value of a Longitudinal Corpus // Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26–28. Lisbon, Portugal. 2004. P. 183–186.

ELAN [Программное обеспечение]. Nijmegen: Max Planck Institute for Psycholinguistics. URL: https://tla.mpi.nl/tools/tla-tools/elan/.

Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus / Ronzhin A. et al. (eds.) SPECOM 2015. Lecture Notes in Artificial Intelligence, LNAI. Vol. 9319. Switzerland : Springer, 2015. P. 268–276.

Sherstinova T. The Structure of the ORD Speech Corpus of Russian Everyday Communication // Text, Speech and Dialogue. Proceedings of the 12th International Conference, TSD-2009. LNAI 5729. Switzerland : Springer, 2009. P. 258–265.

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

RUSSIAN EVERYDAY SPEECH CORPUS "ONE DAY OF SPEECH": CURRENT STATE AND PERSPECTIVES

The article describes some stages of creation and extension of the Russian everyday speech corpus "One Day of Speech" (the ORD corpus), as well as peculiar conditions for publishing its online version. Speech material of the ORD corpus was obtained in natural communicative situations. Volunteer-respondents were people who expressed their willingness to live a day "with a dictaphone dangling around their necks" and fill out several questionnaires. Before recording, the respondents were instructed; they learned to use sound recording equipment and received the "Informant's memo". We asked them to intervene in recording process as little as possible, not to turn off the dictaphone and behave "as usual". All records are anonymous, respondents and their interlocutors are named using codes, personal information in transcripts is marked in a special way. The ORD corpus contains mainly private conversations. An important requirement for the publication of its online version is ensuring anonymity of speakers. Thus, transcripts of sound recordings can be published only after the anonymization of personal names, surnames, and nicknames, and only after any information that may lead to the disclosure of the speaker's identity is excluded from the texts.. The article describes the method proposed for anonymization of personal data and poses the problem of censorship of text transcripts.

Key words: Russian language, everyday spoken speech, speech corpus, personal information, pre-publishing processing.

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. TSD 2009. Ed. by V. Matoušek, P. Mautner. *LNAI*, vol. 57292009. Berlin-Heidelberg, Springer, 2009, pp. 250–257.

Blinova O. V. ["Non-Machine Learning": Experimental Verification of Manual Tokenization Rules]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika-2015"* [Proceedings of International Conference CORPORA 2015]. St. Petersburg, St. Petersburg St. Univ. Publ., 2015, pp. 111–120. (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. *SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI*, vol. 9811. Switzerland, Springer, 2016, pp. 659–666.

Bogdanova-Beglarian N. V. (ed.) *Russkii yazyk povsednevnogo obshcheniya: osobennosti funktsionirovaniya v raznykh sotsial'nykh gruppakh. Kollektivnaya monografiya*

[Everyday Russian Language in Different Social Groups. Collective. Monograph]. St. Petersburg, 2016. 244 p. (In Russ.).

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 russkoi rechi (AR*^{3}*-2017): Trudy sed'mogo mezhdistsiplinarnogo seminara* [Analysis of Spoken Russian Speech (AR^3-2017): Proceedings of the 7th Interdisciplinary Seminar]. St. Petersburg, 2017, pp. 14–20. (In Russ.)

Burnard L. (ed.) Reference Guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, 2007. Available at: http://www.natcorp.ox.ac.uk/docs/URG/ (accessed 23.06.2018).

Campbell N. Speech & Expression; the Value of a Longitudinal Corpus. *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC-2004, May 26–28*. Lisbon, Portugal, 2004, pp. 183–186.

ELAN [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. Available at: https://tla.mpi.nl/tools/tla-tools/elan/ (accessed 23.06.2018).

Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus. Eds. A. Ronzhin et al. *SPECOM 2015, Lecture Notes in Artificial Intelligence, LNAI*, vol. 9319. Springer, Switzerland, 2015, pp. 268–276.

Sherstinova T. The Structure of the ORD Speech Corpus of Russian Everyday Communication. Text, Speech and Dialogue. *Proceedings of the 12th International Conference, TSD-2009. LNAI 5729*. Switzerland, Springer, 2009, pp. 258–265.

Sherstinova T.Ju., Stepanova S. B., Ryko A. I. [Annotation System in the Russian Sound Corpus "One Day of Speech"]. *Materialy XXXVIII Mezhdunarodnoi filologicheskoi konferentsii. Sektsiya: "Formal'nye metody analiza russkoi rechi"* [Proceedings of the XXXVIII International Philological Conference. The Conference Section "Formal Approaches to Analysis of Russian Speech"]. St. Petersburg, St. Petersburg State University Publ., 2009, pp. 66–75. (In Russ.)

Sibata T. [A Twenty-Four-Hour Survey of the Language Life]. *Yazykoznanie v Yaponii* [Linguistics in Japan]. Ed. by V. M. Alpatov, I. F. Vardul'. Moscow, Raduga Publ., 1983, pp. 134–141. (In Russ.)