УДК 81’322.2

МОДЕЛИРОВАНИЕ ПОВСЕДНЕВНОГО РЕЧЕВОГО ПОВЕДЕНИЯ: КОРПУС УСТНОЙ РЕЧИ МОЛОДЕЖИ, ИЛИ ОРД V.2.0

Татьяна Юрьевна Шерстинова
к. филол. н., доцент департамента филологии,
заведующий лабораторией языковой конвергенции
Национальный исследовательский университет
«Высшая школа экономики» – Санкт-Петербург
190068, Санкт-Петербург, наб. канала Грибоедова, 119–121. tsherstinova@hse.ru

Ирина Анатольевна Петрова
стажер-исследователь лаборатории языковой конвергенции
Национальный исследовательский университет
«Высшая школа экономики» – Санкт-Петербург
190068, Санкт-Петербург, наб. канала Грибоедова, 119–121. irina.petrova1011@gmail.com

Для моделирования актуальных речевых процессов повседневной коммуникации необходимы представительные лингвистические ресурсы, такие, например, как корпус «Один речевой день». В докладе представлен новый ресурс, создающийся по методу непрерывной звукозаписи речевого поведения информантов – Корпус устной речи молодежи. Целью создания этого корпуса является получение «свежих» звукозаписей повседневной речи молодежи и студентов – социальных групп, наиболее восприимчивых ко всему новому, что находит свое отражение и на языковом уровне. Создание корпуса не только позволит получить обобщенные статистические характеристики языка современных студентов, но и даст возможность построения формальных моделей, отражающих изменения языковых характеристик в динамике.

Ключевые слова: русский язык; повседневная речь; корпусная лингвистика; социолингвистика; речь молодежи; количественные методы.

Введение

Статья посвящена методике создания мультимедийного языкового ресурса, предназначенного для изучения русского языка повседневного общения на материале живой неподготовленной устной речи, которую мы используем в бытовой и профессиональной коммуникации, и для моделирования речевого поведения человека в естественных ситуациях. Необходимость такого ресурса обусловлена многими факторами, в первую очередь он предназначен

- для изучения языка повседневного общения, лексики, грамматики и прагматики устной речи (как официальной, так и неофициальной);
- для формирования «звучащей памяти» о нашем времени и для сбора коллекции звукозаписей, отражающих «звуковой портрет» эпохи;
- для моделирования речевого общения в разных коммуникативных ситуациях, необходимого как для изучения социальных, психологических и культурных аспектов языка повседневного общения, так и для решения ряда практических задач обучения коммуникативным навыкам людей и роботоподобных систем;
- для настройки и тестирования максимально приближенных к реальности систем синтеза и распознавания речи и других приложений, связанных с разработкой систем искусственного интеллекта, понимающих язык в его естественной форме и умеющих использовать его для звуковой коммуникации.

Материалом для такого ресурса в идеале должны служить записи повседневного речевого общения, выполненные в полевых условиях. В качестве примера таких звуковых коллекций можно назвать демографическую выборку Британского национального корпуса [BNC 2005], японский ESP корпус, являющийся частью большого проекта JST/CREST, посвященного обработке эмоциональной речи [Campbell 2004], а также русский корпус «Один речевой день» (ОРД – https://ord.spbu.ru) – уникальный для рус-

ского языка проект, записи для которого собиралась в 2007–2016 гг. в Санкт-Петербургском государственном университете [Богданова-Бегларян и др. 2019; Asinovsky et al. 2009; Bogdanova-Beglarian et al. 2016; Sherstinova 2009], который стал прототипом для разработки нового ресурса.

При записи ОРД используется методика непрерывной 24-часовой записи, специально разработанная его авторами [Asinovsky et al. 2009]. Добровольцы, желающие принять участие в записи корпуса, получают диктофон, с которым им необходимо прожить свой обычный день. Такой подход позволяет получить образцы речи реальной языковой среды – полевые записи, которые относительно несложно собирать, но довольно сложно обрабатывать.

Связано это с тем, что в полевых звукозаписях естественной речи, по сравнению с «лабораторными» записями, выполненными в профессиональной студии (или даже по телефону), имеют место следующие факторы: 1) наличие большого количества фоновых шумов (в зависимости от конкретной ситуации они могут даже перекрывать речевой сигнал); 2) сильная вариация уровня полезного сигнала при изменении расстояния говорящего от диктофона (например, когда записывающее устройство лежит на столе); 3) неформальный характер бытовых разговоров, как правило, не предназначенных для прослушивания кем-либо иным, кроме самих участников (определенные фрагменты разговора могут быть совершенно непонятны лингвистам, пытающимся транскрибировать и анализировать записанную речь); 4) тематический спектр разговоров, который практически неограничен (здесь могут встречаться как всем понятные «разговоры о погоде», так и какие-то весьма узкие темы, понятные только избранному кругу профессионалов); 5) количество участников разговора, которое может меняться в процессе его развертки и быть фактически неограниченным; 6) наложение речи (то есть одновременное порождение речи несколькими собеседниками) в естественных коммуникативных ситуациях, которое является существенной причиной сложности обработки повседневных полевых разговоров. Кроме того, как и для любого звукового ресурса, отдельную сложность составляет получение транскриптов – расшифровок звучащей речи, позволяющих работать с корпусом как с коллекцией письменных текстов. Комплекс этих факторов объясняет, почему до сих пор наблюдается нехватка представительных звуковых ресурсов, посвященных языку повседневного общения.

Корпус устной речи молодежи – звуковой корпус повседневной русской речи, разработка которого началась на базе Лаборатории языковой конвергенции НИУ ВШЭ в Санкт-Петербурге в 2023 г. По своей сути этот ресурс является идейным и методологическим продолжением корпуса ОРД (см. выше). Необходимость создания нового ресурса устной речи вызвана фактом постоянного изменения повседневной речи прежде всего на лексическом уровне и, главным образом, в речи молодого поколения. Кроме того, с момента последних записей ОРД изменились и технические возможности корпусной лингвистики, в том числе в плане автоматической расшифровки звукозаписей, которые можно задействовать при создании корпуса.

Особенности сбора материала

Хотелось бы отметить некоторые особенности нового ресурса по сравнению с методикой сбора и обработки корпуса ОРД.

1. Кого записываем? Прежде всего, необходимы звукозаписи молодежи и студенчества, так как именно в речи этой возрастной группы можно ожидать новых языковых явлений и неологизмов [Русский язык повседневного общения 2016]. Однако мы принимаем в корпус и речь волонтеров из других возрастных групп. Нет ограничений также на род занятий, место основного проживания участников звукозаписей и их другие социальные характеристики. Такой подход не способствует сиюминутной сбалансированности материала, однако позволит сохранить уникальные записи, полученных от самых разных говорящих (Слово не воробей, вылетит – не поймаешь!), а в дальнейшем даст возможность масштабировать создаваемый ресурс до национального корпуса русской повседневной звучащей речи.

2. Что записываем? Основной единицей звукозаписи, полученной от информантов, остается их «речевой день», т. е. множество звуковых файлов, отражающих их речевую коммуникацию с того момента, как они утром проснулись, до того момента, как они легли спать вечером. Как и при записи ОРД, мы просим добровольцев самих выбирать наиболее подходящие дни для записи, а если в течение дня образовались ситуации, звукозаписи которых по тем или иных причинам участники эксперимента не готовы сдавать в корпус, эти фрагменты удаляются[1]. При сборе данных для нового корпуса мы решили дать возможность записывать не только речевой день целиком, но и отдельные речевые ситуации (например, «отмечание дня рождения» или «беседа с другом»). Это позволяет расширить общее количество речевого материала и разнообразие информантов, представленных в корпусе, что идет на пользу его статистической представительности.

3. Как записываем? Сама методика записи остается практически неизменной. Запись осуществляется на профессиональный цифровой диктофон, который может как висеть на шее информанта при его передвижении в пространстве, так и стационарно лежать на столе при осуществлении записи в помещении. Однако для новых звукозаписей был обновлен набор звукозаписывающих устройств[2]. Формат выходного файла: WAV, стерео, 16 бит, 44 100 Гц. Кроме того, принимая во внимание существенное повышение качество звукозаписи персональной мобильной техники, было принято решение разрешить информантам производить запись и на свои собственные смартфоны. При этом правила о непрерывности звукозаписи остаются неизменными. Тестовые эксперименты показали, что и при записи на современный телефон можно получить речевой материал довольно высокого качества.

Первые звукозаписи для нового корпуса были получены весной 2023 г. На этом этапе исследования к участию были приглашены только студенты ОП «Филология» НИУ ВШЭ Санкт-Петербург для того, чтобы отработать методику сбора материала и протестировать технику. На втором этапе пул кандидатов был расширен за пределы вуза. Единственный критерий отбора добровольцев для записи – их возраст должен быть не менее 18 лет.

Каждый участник, согласившийся на запись, получает набор, в который входят звукозаписывающее устройство, дополнительные комплектующие, памятка информанта. В памятке описана инструкция по использованию диктофона и основные правила ведения записи, например: «Звукозапись выполняется в течение 12–16 часов (с утра до вечера) при минимальном вмешательстве участника эксперимента в процесс звукозаписи. Диктофон останавливается только на время смены источников питания (примерно раз в 9 часов)».

Для полноты данных, каждому из информантов необходимо заполнить социологическую анкету, прототипом которой является опрос, предлагавшийся авторами ОРД. Социологическая анкета предполагает заполнение следующих характеристик: 1) пол; 2) возраст; 3) место рождения; 4) место жительства; 5) опыт проживания в отличном от текущего места жительства регионе; 6) родной язык; 7) другие языки, которыми владеет информант; 8) профессия родителей; 9) уровень образования; 10) полученная специальность; 11) опыт работы, если есть. Новый пункт «профессия родителей» был введен в анкету вместо вопроса «социальное происхождение», который часто вызывал замешательство в ответах информантов ОРД. Помимо информанта, подобные сведения заполняются и для всех основных коммуникантов.

Кроме того, анкета информанта включает согласие на передачу данных в Корпус устной речи молодежи для их использования в научных целях и согласие на обработку персональных данных. Дополнительно в анкету внесен пункт и о согласии/несогласии информантов на открытую публикацию полученного от него речевого материала, при этом возможны следующие варианты ответов: «Да, можно опубликовать запись полностью», «Я буду не против публикации большей части звукозаписи»[3], «Нет, публиковать аудиозапись нельзя». Заполнение анкеты информанта и основных коммуникантов происходит онлайн с помощью облачного сервиса.

Информантам также предлагается пройти психологический тест – пятифакторный опросник личности 5PFQ (https://testometrika.com/personality-and-temper/five-factor-personality-test-ipip-neo-pi-r). С его помощью измеряется степень выраженности каждого из пяти факторов «большой пятерки»: экстраверсии, конформности, сознательности, эмоциональной стабильности, открытости новому опыту [Хромов 2000] (см. Рис.).

Помимо психологического теста, информантам дается тест на определение их пассивного словарного запаса (https://www.myvocab.info), основанный на статистическом подходе Дж. Рида [Read 2000], который предполагает, что «вероятность знания респондентом слов, используемых в языке (в книгах, телепередачах, речи) одинаково часто, примерно одинакова. Это позволяет проверять знание респондентом не всех слов языка, что заведомо невозможно, а только небольшого количества специально отобранных тестовых слов, каждое из которых представляет целую группу слов примерно одинаковой частотности» [Головин 2015: 148].

Во время процесса записи участники исследования должны вести «Дневник речевого дня», в котором кратко отмечаются коммуникативные события (с кем, в какой период времени, при каких обстоятельствах информант общается в течение дня). В настоящее время этот «дневник» реализован посредством Телеграм-бота.

Таким образом, при создании корпуса было решено полностью отказаться от бумажной документации: вся коммуникация с информантами, их анкетирование и тесты осуществляются онлайн.

Запись корпуса осуществляется анонимно. После получения аудиофайлов от информантов, их данные шифруются. Каждому участнику присваивается код вида AF001, где первая буква «A» – знак информанта, вторая буква – обозначение пола «F/M». Далее следует порядковый номер записи.

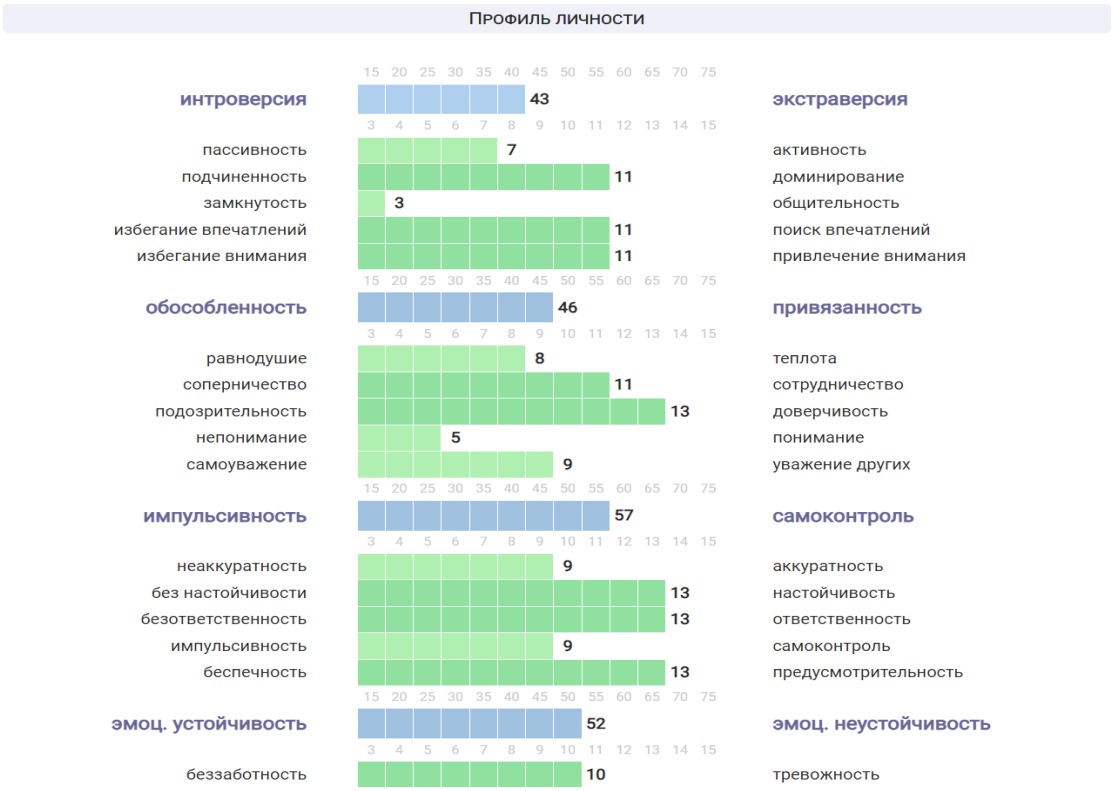

Рисунок. Пример результата прохождения информантом пятифакторного теста (фрагмент)

Принципы обработки речевого материала

Многочасовые звуковые файлы, полученные от информантов, прослушиваются, из них удаляются продолжительные паузы или шумовые фрагменты, не несущие полезной информации. После этого файлы сегментируются ручным образом на макроэпизоды – «крупные эпизоды, объединенные местом коммуникации, ее условиями и участниками (например, “завтрак в кругу семьи”, “дорога на работу”, “рабочее совещание”, “работа с клиентами в офисе”, “покупки в магазине” и т. п.)» [Шерстинова 2013: 450]. Каждый из них аннотируется по методике формального описания макроэпизодов. Как правило, продолжительность макроэпизода составляет от 15 до 40 минут. Каждый эпизод получает кодовое название, в котором отражены название подкорпуса, уникальный код информанта и порядковый номер эпизода (напр., escAF001-01). Формат выходного файла, содержащего преимущественно речь, – моно, 16 бит, 22 050 Гц.

Аннотирование макроэпизодов выполняется по методике ОРД [там же]: выделяется основной тип коммуникации («бытовой разговор», «профессиональная беседа», «публичная речь» и др.), условия коммуникации (например, «телефонный разговор», «застолье», «во время совершения покупок» и др.). Кроме того, тегами помечаются наличие монологической речи, пения, чтения, конфликтных ситуаций и др. особенности коммуникативной ситуации.

Помимо этого, для каждого макроэпизода отмечается социальная роль информанта – родственник (муж, жена, дочь, брат и т. д.), клиент или представитель сервиса, учитель или ученик, сокурсник, коллега; помечается также, когда человек разговаривает сам с собой. Помимо роли информанта, указываются его собеседники и кем они для него являются. Маркируется и место коммуникации: дом, офисное помещение, учебное заведение, поликлиника, улица и т. д.

Звуковые файлы, соответствующие макроэпизодам, являются основной единицей описания в корпусе. Одновременно с их аннотированием, экспертом отмечается качество речевого сигнала с точки зрения зашумленности: 1 – высокое качество записи при минимуме фоновых шумов, позволяющее осуществлять на этом материале фонетические исследования речи; 2 – относительно хорошее качество звукозаписи; 3 – среднее качество записи; 4 – очень зашумленные звукозаписи.

После сегментации звукозаписи передаются на расшифровку для получения транскриптов. При создании корпуса ОРД расшифровки выполнялись экспертами вручную в программе ELAN (http://tla.mpi.nl/tools/tla-tools/elan), при этом для перевода в текст на основных аннотационных уровнях одной минуты речи уходил примерно один час работы эксперта. После этого транскрипт проверялся как минимум дважды с привлечением других экспертов, вносящих свои коррективы в первичную расшифровку. Такой

трудоемкий подход к переводу звукозаписей в текст объясняет тот факт, что к настоящему времени лишь незначительная доля (не более 25%) богатой коллекции ОРД имеет текстовые расшифровки, в то время как большая часть звукозаписей еще ждет своего часа.

Поскольку средства распознавания речи за последнее время стали существенно лучше, при создании нового корпуса повседневной речи было решено привлечь к получению расшифровок современные модели распознавания. На этапе получения транскриптов принято решение отказаться от использования ELAN, а использовать для первичной расшифровки средства автоматической диаризации (разделения речевого потока на сегменты, соответствующие речи отдельных говорящих) и автоматического распознавания речи с последующей ручной коррекцией. Для пилотных экспериментов были использованы две модели – акустическая модель, разработанная компанией НТР (https://ntr.ai), и широкоизвестная многоязычная нейронная сеть Whisper [Radford et al. 2022: электр. ресурс], созданная разработчиком chatGPT компанией Open AI (https://openai.com). Результаты тестовых расшифровок «речевых дней» описаны в работах [Колпащикова 2023; Sherstinova, Kolobov, Mikhaylovskiy 2023].

Текущая статистика собранных данных

На данный момент общий объем полученного от информантов звукового материала равен 377 часам; 113 из них записаны на диктофон, встроенный в мобильный телефон, остальные – на профессиональные звукозаписывающие устройства. Средняя продолжительность записей от одного человека составляет 11 часов, максимальное длительность записи, полученной от одного человека – 62 часа.

Из 377 часов записей прослушаны и обработаны 302. После удаления неречевого материала, объем звуковой части корпуса составляет 190 часов, которые относятся к 552 макроэпизодам. Средняя длительность эпизода – 21 минута. Качество полученного речевого материала отражено в Таблице, которая показывает, что использованная методика действительно дает довольно большой процент «чистых» звукозаписей.

Таблица

Качество звучания макроэпизодов

Оценка	Качество звукозаписи	Количество эпизодов
1	Отлично	256
2	Хорошо	162
3	Удовлетворительно	79
4	Плохо	55

Аудиозаписи получены от 34 человек в возрасте от 18 до 60 лет; из них 31 женщина и 3 мужчины. Гендерный дисбаланс определяется тем, что звукозаписи были начаты в студенческой филологической среде, известной своей гендерной несбалансированностью. Предполагается, что дальнейшее пополнение корпуса выравняет диспропорцию в этом и других социальных аспектах. Средний возраст информантов – 21 год, большинство из них россияне, также есть представители Казахстана, Беларуси и Киргизии. Для половины информантов родной город – Санкт-Петербург. При этом практически все отмечают русский как свой родной язык (за исключением информанта из Беларуси).

Результаты психологического тестирования и теста на пассивный словарный запас получены для 23 информантов. Результаты теста на определение словарного запаса первой группы информантов-филологов находятся в диапазоне от 68 000 до 106 000 слов с медианным значением 85 000.

Что касается дистрибуции коммуникативных ситуаций, с которыми сталкиваются студенты-филологи, то большая часть из них (около 80%), как и в записях корпуса ОРД, относятся к бытовому общению. Около 16% всех эпизодов относится к коммуникации между коллегами. Также присутствует учебная и публичная коммуникация, общение с внешними организациями. Около половины всех эпизодов записаны в домашней обстановке (49%). Хорошо представлена коммуникация на улице (11%), в вузе (9%), офисе (8%), гостях (7%), кафе (7%) и магазине (4%). Примерно 18% всех выделяемых макроэпизодов относятся к общению во время приема пищи.

Что касается социальных ролей, в которых чаще всего выступают студенты-филологи, то они следующие: подруга/друг (38,77%), коллега (9,60%), «герлфренд» (9,24%), сотрудник сервиса (9,24%), разговоры с собой и с домашними животными (7,61%).

Поскольку записи для нового Корпуса только начались, полученную статистику стоит считать предварительной.

Заключение

Разнообразие коммуникативных ситуаций, в которые попадают информанты в процессе звукозаписи своих «речевых дней», и тех социальных ролей, в которых им приходится при этом выступать, дает уникальную возможность использовать полученный таким образом речевой материал для изучения и моделирования речевого поведения жителей Санкт-Петербурга. Планирующееся расширение корпуса с привлечением говорящих из самых разных профессиональных сфер и из разных регионов страны повысит ре-

презентативность корпуса и достоверность выполняемых на его основе теоретических и практических работ. При этом кажется целесообразным активно привлекать краудсорсинговые записи, а также наращивать количество звукозаписей, открытых для свободного использования.

Создание корпуса позволит получить обобщенные статистические характеристики языка современных студентов на разных языковых уровнях от фонетики до прагматики, а сравнение новых звукозаписей с речевым материалом ОРД, полученным 7–15 лет назад, даст возможность построения формальных моделей, отражающих изменения языковых характеристик в динамике.

Одной из целей работы может стать разработка на базе создаваемого ресурса национального корпуса звукозаписей русской повседневной устной речи.

Благодарности

Публикация подготовлена в результате проведения исследования по проекту «Текст как Big Data: моделирование конвергентных процессов в языке и речи цифровыми методами», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2023 г.

Авторы выражают свою признательность всем волонтерам, принявшим участие в звукозаписи корпуса, а также студентам образовательной программы «Филология» НИУ ВШЭ в Санкт-Петербурге, принявшим активное участие в разработке методологии проекта и организации проведения звукозаписи, в частности, Карине Азаревич, Екатерине Ершовой, Анне Журавлевой, Евгении Колпащиковой, Ольге Маркович, Валерии Мелкозеровой, Александре Монастырской, Мадине Мохаммад, Софии Чеповецкой.

Примечания

[1] Опыт показывает, что этой возможностью добровольцы ОРД пользуются нечасто – они чаще ставят диктофон на паузу в ситуациях, записывать которые они не считают желательным.

[2] В настоящее время запись ведется преимущественно на диктофоны Roland R09-HR, Zoom H1n и Tascam DR-05X.

[3] В этом случае предполагается, что информант отмечает эпизоды речевого дня, которые не могут быть опубликованы.

Список литературы

Богданова-Бегларян Н.В. и др. Корпус русского языка повседневного общения «Один речевой день»: текущее состояние и перспективы / Н.В. Богданова-Бегларян, О.В. Блинова, Г.Я. Мартыненко, Т.Ю. Шерстинова // Труды Института русского языка им. В.В. Виноградова. 2019. Т. 21. С. 101–110.

Головин Г.В. Измерение пассивного словарного запаса русского языка // Социо-и психолингвистические исследования. 2015. №. 3. С. 148–159.

Колпащикова Е.О. Писатель Робин Дранаттагор: апробация модели Whisper на русскоязычной звучащей речи // Социо- и психолингвистические исследования. 2023. Вып. 11. (В печати).

Русский язык повседневного общения: особенности функционирования в разных социальных группах / под ред. Н.В. Богдановой-Бегларян // СПб: Лайка, 2016. 244 с.

Хромов А.Б. Пятифакторный опросник личности. Курган: Изд-во Курган. гос. ун-та, 2000. 23 с.

Шерстинова Т.Ю. Коммуникативные макроэпизоды в корпусе повседневной русской речи «Один речевой день»: принципы аннотирования и результаты статистической обработки // Корпусная лингвистика – 2013: Труды Междунар. науч. конф. СПб., 2013. С. 449–456.

Asinovsky A. et al. The ORD Speech Corpus of Russian Everyday Communication “One Speaker’s Day”: Creation Principles and Annotation / A. Asinovsky, N. Bogdanova, M. Rusakova, A. Ryko, S. Stepanova, T. Sherstinova // Lecture Notes in Artificial Intelligence (LNAI). 2009. Vol. 5729. P. 250–257.

BNC – Reference Guide for the British National Corpus. [Электронный ресурс]. URL: http://www.natcorp.ox.ac.uk/docs/URG.xml (дата обращения: 20.11.2023).

Bogdanova-Beglarian N. et al. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech / N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, O. Ermolova, E. Baeva, G. Martynenko, A. Ryko // Lecture Notes in Artificial Intelligence (LNAI). 2016. Vol. 9811. Pp. 659–666.

Campbell N. Speech & Expression; the Value of a Longitudinal Corpus // Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC’04) / eds. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva. Lisbon, 2004. Pp. 183–186.

Radford A. et al. Robust Speech Recognition via Large-Scale Weak Supervision / A. Radford, J.W. Kim, Tao Xu, G. Brockman, C. McLeavey, I. Sutskever. 2022. [Электронный ресурс]. URL: https://cdn.openai.com/papers/whisper.pdf (дата обращения: 11.10.2023).

Read J. Assessing Vocabulary. Cambrige University Press, 2000. 294 p.

Sherstinova T. The Structure of the ORD Speech Corpus of Russian Everyday Communication // Lecture Notes in Artificial Intelligence (LNAI). 2009. Vol. 5729. Pp. 258–265.

Sherstinova T., Kolobov R., Mikhaylovskiy N. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level // Lecture Notes in Computer Science (LNCS). 2023. (в печати).

MODELING EVERYDAY SPEECH BEHAVIOR: A CORPUS OF ORAL SPEECH OF YOUTH, OR ORD V.2.0

Tatiana Y. Sherstinova
Associate Professor, Department of Philology
National Research University Higher School of Economics – St Petersburg

Irina A. Petrova
Research Intern, Linguistic Convergence Laboratory
National Research University Higher School of Economics – St Petersburg

To effectively model contemporary speech processes within daily communication, comprehensive linguistic resources, such as the ORD corpus, are indispensable. This paper introduces a novel resource which was being developed using a continuous audio recording methodology capturing informant's verbal behaviors – youth oral speech corpus named ESC (Everyday Student Conversations) The primary objective behind this corpus' creation is to procure up-to-date recordings of everyday speech from youths and students. These demographic groups are notably receptive to novel influences, which manifest linguistically. The establishment of this corpus not only facilitates a holistic statistical overview of modern students' language but also paves the way for the development of formal models that dynamically reflect linguistic shifts over time.

Keywords. Russian language; everyday speech; corpus linguistics; sociolinguistics; youth discourse; quantitative methods.