Наталья Викторовна Богданова-Бегларян
Татьяна Юрьевна Шерстинова
Александр Семенович Асиновский
Ольга Владимировна Блинова
Елена Валерьевна Маркасова
Анастасия Игоревна Рыко
Россия, Санкт-Петербургский государственный университет

Звуковой корпус русского языка: новая методология анализа устной речи[1]

Ключевые слова: звуковой корпус, спонтанная речь, повседневная коммуникация, лексикографическая база данных, монологическая речь, диалогическая речь
Key words: speech corpus, spontaneous speech, everyday communication, lexicographical database, monologic speech, dialogic speech

Abstract

The article describes the Speech Corpus of the Russian Language, created at the Faculty of Philology of St. Petersburg State University. The corpus consists of two modules. The first module contains monologic speech of native speakers, recorded in the situations of various speech scenarios (reading, retelling, picture description, story-telling), strictly balanced along linguistic, social and psychological criteria. The second module contains everyday speech, mostly dialogic, recorded using the method of 24-hour continuous recording ("One Speaker's Day"). The Speech Corpus of the Russian Language contains more than 500 hours of recordings; the specialized linguistic information system is built upon the corpus content, and numerous multidimensional studies of Russian oral communication are currently being conducted.

Звуковой корпус русского языка (ЗКРЯ) мыслится его создателями как сложная структура, позволяющая осуществлять мониторинг современной повседневной речи, фиксировать язык в его наиболее естественной форме, а также

[1] Исследование выполнено в рамках тематического плана фундаментальных НИР СПбГУ «Интегральное моделирование звуковой формы естественных языков: коммуникативно-семиотический аспект» (шифр проекта 31.0.23.2010).

получать материал для многоуровневого описания русской звучащей речи, для решения целого ряда прикладных лингвистических задач и для преподавания языка в его звуковой форме. Корпус включает в себя специально созданную информационную среду и программный инструментарий для нужд интегрального моделирования естественной речи.

Работа над созданием Звукового корпуса началась на филологическом факультете СПбГУ еще в 2007 г. и сразу проводилась по двум направлениям, акцентирующим исследовательское внимание на двух разных типах повседневной речи. В результате ЗКРЯ включает два блока (модуля), устроенных принципиально по-разному, но преследующих общую цель – фиксацию состояния современной русской речи во всем ее жанровом и тематическом разнообразии, в неразрывной связи как с ситуацией общения, так и с говорящим индивидом и его социальными отношениями с собеседниками (коммуникантами).

Первый блок корпуса – «сбалансированная аннотированная текстотека» (САТ) – изначально достаточно строго сбалансирован по разным параметрам – социологически, психологически и собственно лингвистически, в нем реализован *принцип ковчега* («каждой твари по паре»).

Лингвистическая балансировка материала заключается в том, что все тексты построены в рамках комплекса *коммуникативных сценариев*: *чтение* и *пересказ* (сюжетный и несюжетный исходный текст); *описание изображения* (сюжетное и несюжетное); *свободный рассказ на заданную тему* (знакомую и незнакомую).

Социолингвистическая балансировка материала предполагает учет социальных характеристик информантов, таких как пол, возраст, профессиональная принадлежность, профессиональное или непрофессиональное отношение к речи, уровень речевой компетенции и некоторые другие.

Психолингвистическая балансировка материала предполагает учет психологических характеристик информантов, прежде всего их экстравертности/интровертности (в основе дифференциации – психологический тест Г. Айзенка).

К настоящему времени данный модуль содержит монологи, 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 два блока интерферированной русской речи иностранцев: американцев и китайцев[2]. Всего это около 700 текстов и около 40 часов звучания.

Вторым блоком Звукового корпуса является модуль «Один речевой день» (ОРД), который ставит своей целью изучение речевого поведения

[2] В настоящее время эта часть САТ пополняется записями русской речи французов.

носителя языка в течение дня (с использованием методики 24-часовой записи[3]) [Asinovsky и др. 2009].

Приоритетная задача создания данного корпуса заключается в том, чтобы получить записи русской спонтанной речи в максимально *естественных* условиях [см. об этом подробнее: Степанова и др. 2008].

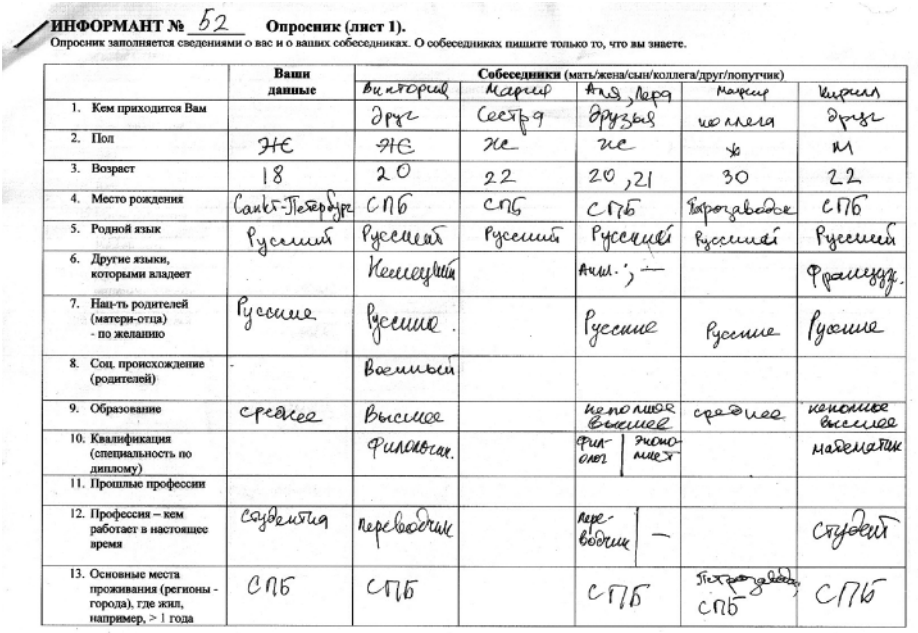
Пилотная звукозапись 2007 г. проводилась по *принципу невода*: в среду носителей языка забрасывалась широкая сеть, вытягивая все, что в нее попало; полученный речевой материал становится объектом многоуровневого исследования. Хотя на первых этапах записи материала выборка информантов не была полностью сбалансирована, она до некоторой степени отражает социальный и психологический срез современного общества. «Язык есть кусочек жизни людей» [Щерба 1974: 98], и используемая методика звукозаписи позволяет увидеть именно эту реальную, естественную, а не искусственно созданную в лабораторных условиях, жизнь, отраженную в речи[4].

Запись модуля ОРД проводилась с использованием диктофона, закрепленного на информанте стационарно, в течение целого дня (иногда – нескольких дней). Каждый информант должен был вести своеобразный дневник «речевого дня», указывая в нем своих коммуникантов, а также ситуацию, в которой происходила коммуникация (например, «в магазине», «в метро», «общение с друзьями» и т. п.). Кроме того, все информанты, участвующие в записи, заполняли социологическую анкету и проходили психологическое тестирование (тесты Г. Айзенка, Кеттела и FPI), что открывает новые возможности исследования материала – с учетом психо-социальных характеристик говорящего и его социальной роли в конкретном коммуникативном акте. На рисунках 1–2 представлены образцы некоторых документов, легших в основу базы данных ОРД: социологическая анкета информанта, описание его коммуникантов и дневник речевого дня.

---

[3] Данный метод был разработан в рамках японской школы языкового существования [см., напр.: Сибата 1983] еще в середине XIX в. Подобный метод звукозаписи был использован в 1990-х гг. прошлого века в Великобритании при сборе материала для устной части Британского Национального корпуса [см.: British National Corpus 2007]. Один из крупных проектов последнего времени – высокотехнологичный корпус японской спонтанной речи JST/CREST ESP Project, задачей которого является описание реальной спонтанной речи японцев для обучения этой речи многофункциональных роботов. Для русского языка такой метод используется впервые.

[4] Надо сказать, что к моменту написания этой статьи (осень 2014 г.) данный принцип претерпел некоторые изменения: на новом этапе формирования корпуса мы уже пытаемся достичь известного баланса в составе информантов ОРД: гендерного, возрастного, образовательного и в некоторых других социальных аспектах [см. об этом подробнее: Баева 2014]. Однако если помнить, что Звуковой корпус содержит речь не только самих информантов, но и в значительно большей доле речь их коммуникантов, состав которых никакой балансировке не поддается, можно все же считать *принцип невода* основным при формировании данного модуля ЗКРЯ.

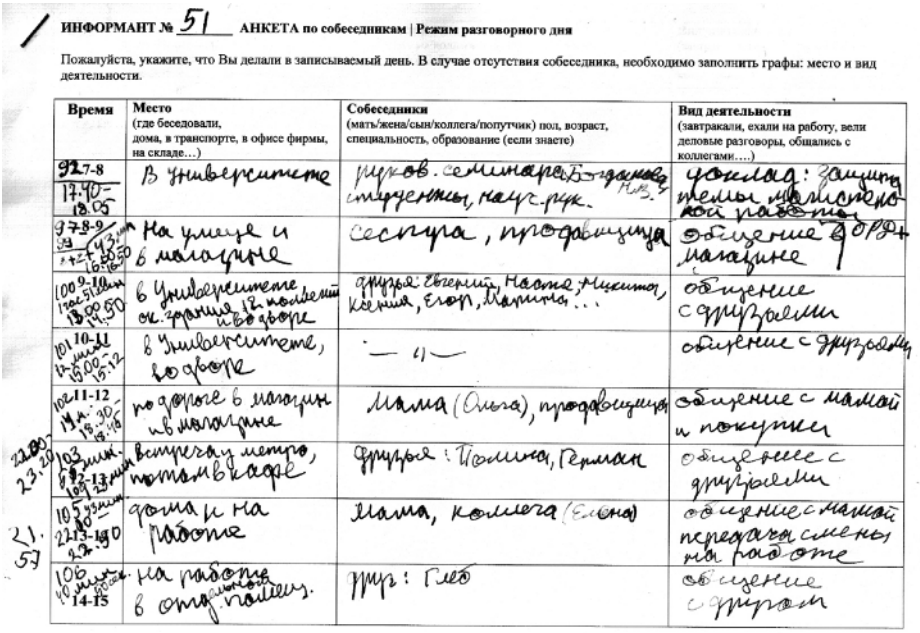


ИНФОРМАНТ № 52 **Опросник (лист 1).**

Опросник заполняется сведениями о вас и о ваших собеседниках. О собеседниках пишите только то, что вы знаете.

	Ваши данные	Собеседники (мать/жена/сын/коллега/друг/попутчик)				
		Виктория	Мария	Анд, Лера	Мария	Кирилл
1. Кем приходится Вам		друг	сестра	друзья	коллега	друг
2. Пол	ЖС	ЖС	ЖС	ЖС	Ж	М
3. Возраст	18	20	22	20, 21	30	22
4. Место рождения	Санкт-Петербург	СПБ	СПБ	СПБ	Екатеринбург	СПБ
5. Родной язык	Русский	Русский	Русский	Русский	Русский	Русский
6. Другие языки, которыми владеет		Немецкий		Англ.; —		Французск.
7. Нац-ть родителей (матери-отца) - по желанию	Русские	Русские		Русские	Русские	Русские
8. Соц. происхождение (родителей)	-	Военный				
9. Образование	среднее	Высшее		неполное высшее	среднее	неполное высшее
10. Квалификация (специальность по диплому)		Филология		Филолог / экономист		математик
11. Прошлые профессии						
12. Профессия – кем работает в настоящее время	студентка	переводчик		переводчик / —		студент
13. Основные места проживания (регионы - города), где жил, например, > 1 года	СПБ	СПБ		СПБ	Екатеринбург, СПБ	СПБ

Рис. 1. Социологическая анкета информанта № 52 (И52)



ИНФОРМАНТ № 51 **АНКЕТА по собеседникам | Режим разговорного дня**

Пожалуйста, укажите, что Вы делали в записываемый день. В случае отсутствия собеседника, необходимо заполнить графы: место и вид деятельности.

Время	Место (где беседовали, дома, в транспорте, в офисе фирмы, на складе…)	Собеседники (мать/жена/сын/коллега/попутчик) пол, возраст, специальность, образование (если знаете)	Вид деятельности (завтракали, ехали на работу, вели деловые разговоры, общались с коллегами….)
92 7-8 17.40–18.05	В университете	преп. семинара Богданова, студентки, науч. рук.	доклад: защита темы магистерской работы
97 8-9 16.50	На улице и в магазине	сестра, продавщица	общение с сестрой в магазине
100 9-10 13.00 50	в Университете, ожидание у входа	друзья: Евгений, Настя, Никита, Ксения, Егор, Марина…	общение с друзьями
101 10-11 15.00–15.12	в Университете, во дворе	— // —	общение с друзьями
102 11-12 19.30–19.40	по дороге в магазин, в магазине	мама (Ольга), продавщица	общение с мамой и покупки
103 12-13	встреча у метро, потом в кафе	друзья: Полина, Герман	общение с друзьями
105 13-14	дома и на работе	мама, коллега (Елена)	общение с мамой, передача смены на работе
106 14-15	на работе в отделении	друг: Глеб	общение с другом

Рис. 2. Дневник речевого дня информанта № 51 (И51)

К лету 2014 г. блок ОРД содержал записи речи 50 информантов (24 мужчин и 26 женщин в возрасте от 17 до 70 лет), а также примерно 650 их собеседников (возраст 3–85 лет), всего около 400 часов звучания[5].

Расшифровка и аннотирование звукозаписей корпуса ОРД выполняются в среде ELAN («EUDICO Linguistic Annotator»), разработанной в Институте психолингвистики Макса Планка для аннотирования мультимедийного контента (http://www.lat-mpi.eu/tools/elan). ELAN поддерживает:

– неограниченное количество задаваемых пользователем уровней аннотации (Tiers);
– визуализацию аудио- и/или видеосигналов одновременно с полученными аннотациями;
– временну́ю привязку аннотаций к мультимедийному потоку;
– сложные связи аннотаций друг с другом;
– различные шрифты и кодировки;
– экспорт данных в виде текстовых файлов табличного вида (*tab-delimited text*);
– импорт и экспорт между ELAN, PRAAT, ToolBox и другими популярными лингвистическими программами;
– поисковые опции [Hellwig и др. 2014].

Текст *расшифровок* записывается в стандартной орфографии, с соблюдением некоторых разработанных авторами правил. Так, начало реплики не выделяется заглавной буквой (такие буквы используются только в именах собственных), интонационное членение звукового потока показывается с помощью маркеров синтагматического и фразового членения (/ и //), введены обозначения пауз разного типа:

*П – пауза;
() – краткая пауза хезитации;
(…) – длительная пауза хезитации;
(*а-а*), (*м-м*), (э-э) – заполненная пауза хезитации.

Используются также некоторые дополнительные символы:

*Н – неразборчивый фрагмент речи;
Торфяновка(?) – вероятная/спорная расшифровка;
по… – оборванное слово;
– смена говорящего, @ – наложение речи разных говорящих;
*С – смех, *К – кашель, *В – вдох/вздох; и др. (подробнее о правилах расшифровки материала ОРД см.: [Шерстинова и др. 2009]).

Первичное аннотирование звучащего материала выполняется в программе ELAN для следующих уровней:

[5] Запись материала для блока ОРД сейчас активно продолжается. К моменту написания этой статьи он уже включает речь более 110 информантов, социальные характеристики которых см.: [Баева 2014].

– Phrase (реплики говорящих);
– Speaker (код говорящего);
– Events (невербальные аудиособытия);
– Voice (качество голоса говорящего);
– PhonetCom (фонетический комментарий);
– PhraseComment (фразовый комментарий);
– Notes (общий комментарий);
– Episode (мини-эпизод речевой коммуникации) (см. пример на рис. 3).

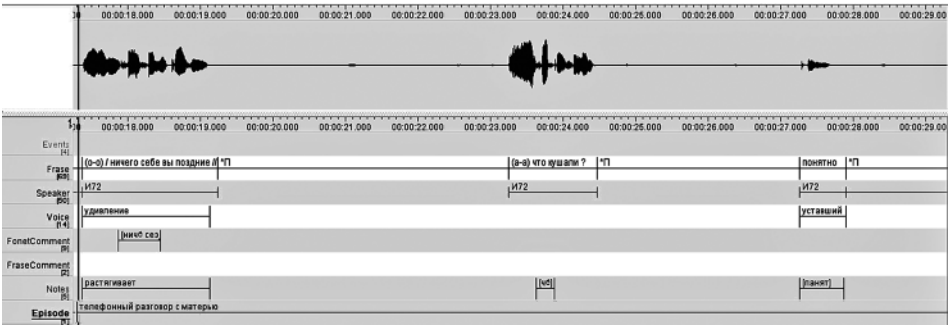


Рис. 3. Пример многоуровневого аннотирования в среде ELAN

Уже первые наблюдения над записанным корпусным материалом показали, что в течение «речевого дня» женщины говорят в целом больше, чем мужчины (см. рис. 4): в частности, от наиболее молчаливого мужчины (И17) (вместе с его коммуникантами) получено за день менее двух часов речевой продукции, а от наиболее разговорчивой женщины (И12) (также вместе с ее коммуникантами) – более 16 часов.

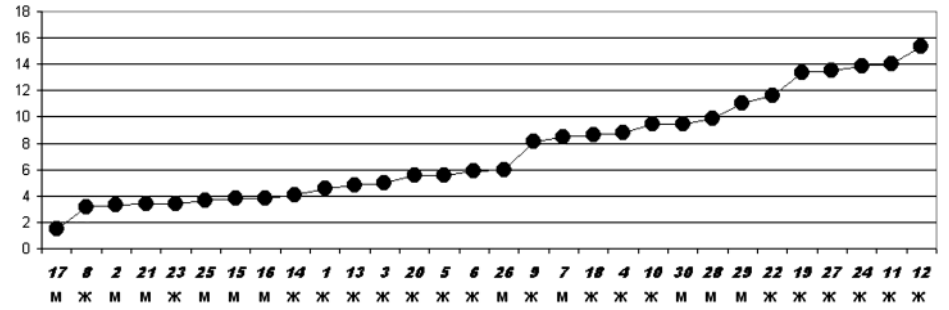


Рис. 4. Общее количество часов записи для первых 30 информантов

Оказалось, что больше всего мы говорим днем (56% всей речевой продукции за день), а вечером – в два раза больше, чем утром (29 *vs*. 15%) (см. рис. 5).

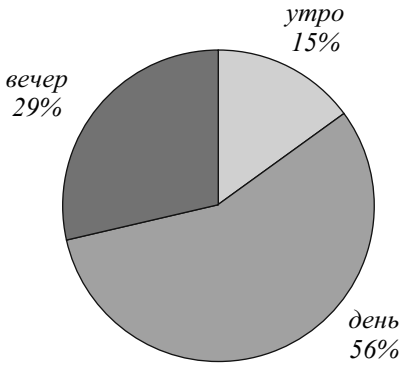


Рис. 5. Распределение речевого материала по времени суток

Максимальное количество разговоров осуществляется «дома» (точнее, в месте постоянного проживания) и на работе (в офисе). Кроме того, в корпусе ОРД хорошо представлена коммуникация в учебных заведениях, «в гостях» и сервисных службах, а также в медицинских центрах и на улице (см. табл. 1) [Шерстинова 2013].

Место коммуникации	**%**	**Место коммуникации**	**%**
дом	18,0	др. обществ. место	5,0
офис	16,0	кафе	5,0
учебное заведение	12,0	транспорт	5,0
в гостях	8,0	магазин	4,0
сервис-центр	7,0	казарма	2,0
на улице	6,0	подсобное помещение	1,0
поликлиника	6,0	прочее	5,0

Табл. 1. Обобщенное распределение речевого материала по «месту действия»

Различаются и обобщенные «речевые дни» мужчин и женщин: первые чаще общаются на разного рода мероприятиях и по дороге – на работу или на те же мероприятия. Максимум женской речевой активности приходится на домашние разговоры утром и вечером (см. табл. 2) [Шерстинова 2008].

Помимо этих общих наблюдений над спецификой нашего повседневного общения, получен и ряд конкретных данных, связанных с разными уровнями обработки звучащего материала.

Так, удалось определить средний *темп речи* наших информантов: 5,31 слога в секунду (сл/с); разброс по разным информантам – от минимального

3,6 сл/с до максимального 6,7 сл/с, что в целом выше, чем, например, в норвежском (3,5–4,5 сл/с), в северном стандартном голландском (5,2 сл/с) или во французском языке (по некоторым данным – 4,31 сл/с), но существенно

Эпизод	**«Женский» день**		**«Мужской» день**		**Разность, %**
	(*мин.*)	%	(*мин.*)	%	**«Ж» – «М»**
завтрак	90	0,94	6	0,16	0,79
дом. разговоры/утро	598	6,27	126	3,26	3,01
в гостях у друзей/утро	16	0,17	0	0,00	0,17
работа дома за компьютером	195	2,05	0	0,00	2,05
дорога на работу/ мероприятие	703	7,38	472	12,22	-4,85
работа/учеба	3973	41,69	1643	42,55	-0,87
обед/ланч	311	3,26	117	3,03	0,23
застолье на работе	0	0	57	1,48	-1,48
сервис-службы/ госучреждения	24	0,25	55	1,42	-1,17
покупки/улица	135	1,42	32	0,83	0,59
прогулка	96	1,01	100	2,59	-1,58
поликлиника/врачи	116	1,22	35	0,91	0,31
хобби-спорт	10	0,10	0	0,00	0,10
мероприятия	60	0,63	360	9,32	-8,69
дома днем	8	0,08	100	2,59	-2,51
на даче/в доме	114	1,20	0	0,00	1,20
вечеринка в кафе/в гостях	991	10,40	321	8,31	2,08
в гостях	320	3,36	140	3,63	-0,27
дорога домой	349	3,66	104	2,69	0,97
ужин	185	1,94	0	0,00	1,94
дом. разговоры/вечер	1136	11,92	193	5,00	6,92
гости дома/вечер	101	1,06	0	0,00	1,06

Табл. 2. «Женский» речевой день vs. «Мужской» речевой день

ниже, чем в испанском (7,81 сл/с) или в бразильском португальском (6,57 сл/с). Примерно в нашем темпе говорят на английском языке: как в Великобритании (3,16–5,33 сл/с), так и в США (3,1–5,4 сл/с) [подробнее см.: Stepanova 2013].

Установлена и зависимость темпа нашей повседневной речи от различных факторов: гендера (мужчины говорят быстрее женщин), возраста (с возрастом мы говорим медленнее), уровня речевой компетенции (чем выше этот уровень у конкретного говорящего, тем ниже темп его речи; см. рис. 6) и социальной роли коммуникантов (с друзьями мы говорим быстрее, чем с коллегами по работе) [подробнее см.: Stepanova 2013; см. также: Метлова 2014].

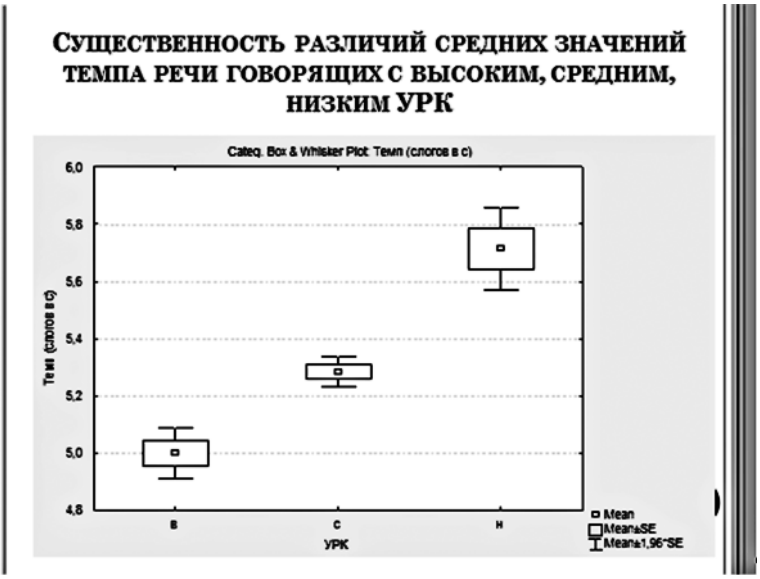


Рис. 6. Зависимость темпа речи от уровня речевой компетенции говорящего

На материале Звукового корпуса уже началась работа по реализации целого ряда инновационных лексикографических проектов, которые в совокупности позволят создать информационно-исследовательскую систему «Язык мегаполиса» (см. рис. 7), предназначенную для создания настраиваемых мультимедийных (интерактивных) словарей, описывающих современную русскую повседневную коммуникацию, а также для многоуровневой обработки словарных данных по запросу пользователя.

Реализация этих проектов имеет важное значение для решения как фундаментальной научно-исследовательской задачи лексикографического описания современного русского языка повседневного общения[6], так и актуальных при-

[6] Ср.: «Именно совокупность лексикографических изданий отражает языковой облик эпохи, представляет наиболее существенные изменения в лексиконе современной языковой личности» [Черняк 2004: 131].

Рис. 7. Возможности лексикографического описания повседневной разговорной речи

кладных задач в области речевых технологий, телекоммуникации, психологии, образования и лингвистического мониторинга современного российского общества. Ниже приводятся некоторые образцы текстов из ОРД, в которых выделены единицы, однозначно требующие дефиниций/описания в одном из предлагаемых словарей:

– *ну то есть имеет смысл* / *да* / ***как бы*** *не… @ просто я … @ не …* ***не гонево*** / *тебе(:)* / *больше не кажется* / *что это всё обман*;
– *и только после того как Глухарева* ***отсмотрит*** / ***отфотографирует*** / *и (…) отправить* / *но вы всё это отложите уже вот*;
– *Ириш* / **П я работала* / **П и () в общем в итоге это в такой* / *вы… вы… вы… вы…* ***вылилось в такой гадюшник*** *что* / ***ну его*** / **П не хочу*;
– *так* / *заставка* // *вот заставка* // *а что ж это у меня получилось-то? *П чистый лист* / ***драсьте пожалста!***
– *извини меня* / ***хахаль*** *из двухкомнатной* ***хаты*** / *@ угу* // **П *Н *П* ***как его ?*** **П там евроремонтик у него* / *пластиковые окна* / *кухонька новая*;
– *ну вот* // **П и тут звонок в дверь* // *стоит этот мужик* // **П* ***типа того что блин*** / **П *Х *П давайте общаться !*
– *у неё … # а я* ***и то и другое*** *(э-э) то есть … # вы с ней очень осторожно*.

Материалы Звукового корпуса, в частности блока ОРД, позволили также уточнить частотность употребления тех или иных частей речи и различных классов словоформ на достаточно представительном материале спонтанной русской речи. Так, в таблице 3 можно видеть распределение словоупотреблений по частям речи в ОРД (устная речь) в сравнении с данными Национального корпуса русского языка (НКРЯ) (основной подкорпус; письменная речь).

Видно, что в повседневной устной и в письменной формах нашей речи практически одинаково часто используются глаголы – по 17%. Различие заключается лишь в том, что в ОРД эта категория слов занимает первое место по частотности, чуть-чуть «опережая» существительное, местоимение-существительное

Часть речи	**НКРЯ**		**ОРД**	
	Абс. кол-во	Отн. кол-во (%)	Абс. кол-во	Отн. кол-во (%)
Существительное	16 93 312	28,7	28 910	15,8
Глагол	10 07 618	17,1	31 449	17,2
Предлог	621 883	10,6	14 486	7,9
Прилагательное	506 851	8,6	6 341	3,5
Союз	471 309	8,0	8 133	4,5
Местоимение-существительное	467 455	8,0	28 369	15,5
Местоимение-прилагательное	277 634	4,7	6 564	3,6
Частица	268 104	4,6	28 646	15,7
Наречие	246 213	4,2	13 797	7,6
Местоимение-наречие	129 369	2,2	7 399	4,0
Местоимение-числительное	—	—	348	0,2
Числительное	102 039	1,7	3 855	2,1
Предикатив	42 280	0,7	1 763	1,0
Вводное слово	25 891	0,4	652	0,4
Числительное-прилагательное	24 535	0,4	852	0,5
Междометие	8 375	0,1	1 296	0,7
ВСЕГО	**5 892 868**	100,0	**182 512**	100,0

Табл. 3. Распределение словоупотреблений по частям речи в НКРЯ и ОРД

и частицу (по 16%), в то время как в письменной речи существительных оказалось в полтора раза больше, чем глаголов – 28%. Тот факт, что существительные в разговорной речи менее употребительны, чем в других сферах языка, отмечают все исследователи. Е. А. Земская объясняет это тем, «что существительные нередко заменяются местоимениями или вообще отсутствуют в тексте, подвергаясь конситуативному и контекстному эллипсису» [Земская 1983: 139]. И действительно, местоимения-существительные в нашем

материале встречаются в два раза чаще (в процентном отношении), чем в письменных текстах НКРЯ: 15,5 *vs*. 8% [подробнее об этом см.: Степанова, Сковпень 2011].

Звуковой корпус русского языка, в первую очередь модуль ОРД, продолжает пополняться новыми записями, но уже и сейчас на этом материале получено множество интересных наблюдений, ставших результатом многоаспектных исследований.

Так, помимо уже названного, в *общетеоретическом плане* разрабатывается само понятие спонтанности речи, происходит поиск единиц описания устного текста, способов его сокращения и приращения, установлена и экспериментально подтверждена градация степеней естественности устной речи.

На *фонетическом уровне* анализируются паузы хезитации, редуцированные формы сверхчастотных слов, фонетические ошибки чтения и говорения. Интересной представляется выявленная тенденция к изохронности структурных единиц звучащей речи [Шерстинова 2010], создаваемая во многом с помощью асемантических дискурсивных ритмообразующих единиц [Богданова-Бегларян и др. 2013]:

– *девять тысяч* ***там*** | *с копейками*;
– ***там*** *чтобы* | *не воровали* | *ничего*;
– ***вот*** *мне* ***там в этом*** | *() пенсионном отделе* | / *одна* ***вот*** *женщина* | *говорит*;
– ***ну*** *все* ***короче*** | *книги у него* | *заканчиваются* | *трагично* (см. рис. 8).

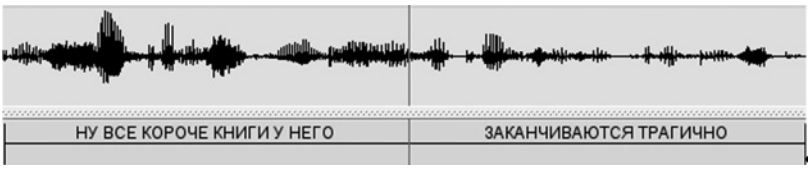
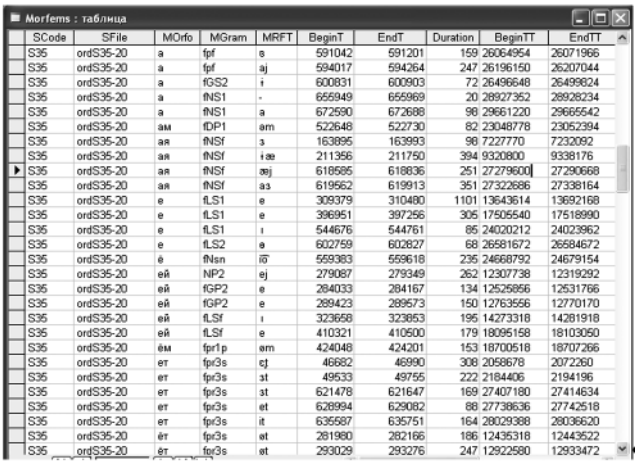


Рис. 8. Осциллограмма и две/четыре изохронные части фразы ***ну*** *все* ***короче*** | *книги у него* | *заканчиваются* | *трагично*

На *лексическом уровне*, в числе прочего, в центре внимания исследователей оказываются «новые» (то есть не попадавшие ранее в зону внимания лексикографов) слова, значения, коннотации, идиомы; строятся конкордансы, частотные словари, ориентированные на тип речи или различные характеристики говорящих, анализируются проявления внутриязыковой интерференции (результат контакта литературного языка с профессиональной речью, просторечием или арго).

Морфологический уровень анализа, помимо дистрибуции частей речи и грамматических классов слов, представлен описанием функционирования глагольных форм, номинативной лексики в текстах разного типа, поиском путей снятия омонимии при анализе звукового потока и т. п. См., например,

на рисунке 9 реализацию грамматических аффиксов в материале ОРД [подробнее см.: Степанова и др. 2010].



Morfems : таблица

SCode	SFile	MOrfo	MGram	MRFT	BeginT	EndT	Duration	BeginTT	EndTT
S35	ordS35-20	а	fpf	ɐ	591042	591201	159	26064954	26071966
S35	ordS35-20	а	fpf	aj	594017	594264	247	26196150	26207044
S35	ordS35-20	а	fGS2	i	600831	600903	72	26496648	26499824
S35	ordS35-20	а	fNS1	-	655949	655969	20	28927352	28928234
S35	ordS35-20	а	fNS1	a	672590	672688	98	29661220	29665542
S35	ordS35-20	ам	fDP1	əm	522648	522730	82	23048778	23052394
S35	ordS35-20	ая	fNSf	ɜ	163895	163993	98	7227770	7232092
S35	ordS35-20	ая	fNSf	iæ	211356	211750	394	9320800	9338176
S35	ordS35-20	ая	fNSf	æj	618585	618836	251	27279600	27290668
S35	ordS35-20	ая	fNSf	aɜ	619562	619913	351	27322686	27338164
S35	ordS35-20	е	fLS1	e	309379	310480	1101	13643614	13692168
S35	ordS35-20	е	fLS1	e	396951	397256	305	17505540	17518990
S35	ordS35-20	е	fLS1	ɪ	544676	544761	85	24020212	24023962
S35	ordS35-20	е	fLS2	ə	602759	602827	68	26581672	26584672
S35	ordS35-20	ё	fNsn	ɪõ	559383	559618	235	24668792	24679154
S35	ordS35-20	ей	NP2	ej	279087	279349	262	12307738	12319292
S35	ordS35-20	ей	fGP2	e	284033	284167	134	12525856	12531766
S35	ordS35-20	ей	fGP2	e	289423	289573	150	12763556	12770170
S35	ordS35-20	ей	fLSf	ɪ	323658	323853	195	14273318	14281918
S35	ordS35-20	ей	fLSf	e	410321	410500	179	18095158	18103060
S35	ordS35-20	ём	fpr1p	əm	424048	424201	153	18700518	18707266
S35	ordS35-20	ет	fpr3s	cʲ	46682	46990	308	2058678	2072260
S35	ordS35-20	ет	fpr3s	ɜt	49533	49755	222	2184406	2194196
S35	ordS35-20	ет	fpr3s	ɜt	621478	621647	169	27407180	27414634
S35	ordS35-20	ет	fpr3s	et	628994	629082	88	27738636	27742518
S35	ordS35-20	ет	fpr3s	it	635587	635751	164	28029388	28036620
S35	ordS35-20	ёт	fpr3s	ət	281980	282166	186	12435318	12443522
S35	ordS35-20	ёт	fpr3s	ət	293029	293276	247	12922580	12933472

Рис. 9. Реализация грамматических аффиксов в материале ОРД

На *синтаксическом уровне* анализируются вставные конструкции разного типа, способы передачи чужой речи, синтаксические трансформации исходных текстов при пересказе, структура предикативных единиц и многое другое.

Дискурсивный уровень анализа представлен исследованием различных черт, свойственных именно устному неподготовленному дискурсу: повторы, перебивы, самокоррекция, коммуникативные установки и стратегии говорящего в разных типах текстов, метакоммуникация. Особенное внимание уделено в наших исследованиях *вербальным хезитативам* и прочим *дискурсивным единицам*, выполняющим в нашей речи множество самых разных функций и формирующим в конечном счете довольно обширный класс *прагматем*.

Помимо анализа прагматем, к *прагматическому уровню* анализа материалов Звукового корпуса можно отнести также разработку методических приемов использования этих материалов в практике преподавания русского языка как иностранного или описание идиолекта (речи конкретной языковой личности в разных коммуникативных условиях).

Наконец, анализируются на нашем материале функции смеха и вздохов в спонтанной речи, являющиеся *паралингвистическими* элементами звукового потока [об этом и многом другом см. подробнее: Богданова-Бегларян 2013; 2014а; 2014б].

В заключение можно сказать, что корпусный подход к анализу устной речи позволяет поставить задачу пересмотра на этом материале практически *всех* накопленных к настоящему моменту лингвистических сведений:

– на лексическом уровне – новые словари;

– на морфологическом уровне – новая грамматика;
– на фонетическом уровне – иной, чем в кодифицированном языке, звуковой облик значимых единиц;
– на словообразовательном уровне – иные модели образования значимых единиц; и пр.

Уже первые наблюдения над материалом показали, что проверка на корпусе самых, казалось бы, очевидных и общепринятых лингвистических утверждений иногда приводит к совершенно неожиданным и интересным результатам.

Литература

Баева Е. М., 2014, О способах социолингвистической балансировки устного корпуса (на примере «Одного речевого дня») [в:] *Вестник Пермского университета. Российская и зарубежная филология*, вып. 4 (28) [в печати].

Богданова-Бегларян Н. В. (отв. ред.), 2013, *Звуковой корпус как материал для анализа русской речи: Коллективная монография*, ч. 1: *Чтение. Пересказ. Описание*, Санкт-Петербург: Филологический факультет СПбГУ.

Богданова-Бегларян Н. В. (отв. ред.), 2014а, *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*, Санкт-Петербург: Филологический факультет СПбГУ.

Богданова-Бегларян Н. В. (отв. ред.), 2014б, *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*, Санкт-Петербург: Филологический факультет СПбГУ (в печати).

Богданова-Бегларян Н. В., Шерстинова Т. Ю., Кислощук А. И., 2013, О ритмообразующей функции дискурсивных единиц [в:] *Вестник Пермского университета. Российская и зарубежная филология*, вып. 2 (22), с. 7–17.

Земская Е. А., 1983, Морфология [в:] Е. А. Земская (ред.), *Русская разговорная речь: Фонетика, морфология, лексика, жест*, Москва: Наука, с. 80–141.

Метлова В. А., 2014, Темп речи в свободной коммуникации: Социолингвистический аспект [в:] *Вестник Пермского университета. Российская и зарубежная филология*, вып. 4 (28) [в печати].

Сибата Т., 1983, Исследование языкового существования в течение 24 часов [в:] В. М. Алпатов (ред.), *Языкознание в Японии*, Москва: Радуга, с. 134–141.

Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю., 2008, Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние формирования [в:] *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»*, вып. 7 (14), Москва: РГГУ, с. 488–494.

Степанова С. Б., Асиновский А. С., Рыко А. И., Шерстинова Т. Ю., 2010, Звуковая реальность словоизменительных аффиксов (по данным Звукового корпуса русского языка) [в:] *Компьютерная лингвистика и интеллектуальные технологии:*

По материалам ежегодной Международной конференции «Диалог», вып. 9 (16), Москва: РГГУ, с. 491–498.

Степанова С. Б., Сковпень О. П., 2011, Дистрибуция частей речи в устной спонтанной речи (на материале Звукового корпуса русского языка «Один речевой день») [в:] А. С. Асиновский (отв. ред.), Н. В. Богданова (науч. ред.), *Материалы XL международной филологической конференции*, вып. 24: *Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков*, Санкт-Петербург: Филологический факультет СПбГУ, с. 200–212.

Черняк В. Д., 2004, Современная лексикография как зеркало социальной жизни современной России [в:] С. И. Богданов, Л. А. Вербицкая, Л. В. Московкин, Е. Е. Юрков (ред.), *Современная русская речь: Состояние и функционирование. Сборник аналитических материалов*, Санкт-Петербург: Филологический факультет СПбГУ, с. 131–155.

Шерстинова Т. Ю., 2008, «Один речевой день» на временной шкале: О перспективах исследования динамических процессов на материале звукового корпуса [в:] *Вестник Санкт-Петербургского университета. Филология. Востоковедение. Журналистика*, сер. 9, вып. 4, ч. 2, с. 227–235.

Шерстинова Т. Ю., 2010, Об изохронности структурных единиц в спонтанной речи (к постановке проблемы) [в:] А. С. Асиновский (отв. ред.), Н. В. Богданова (науч. ред.), *Материалы XXXIX Международной филолологической конференции*, вып. 23: *Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков*, Санкт-Петербург: Филологический факультет СПбГУ, с. 109–118.

Шерстинова Т. Ю., 2013, Коммуникативные макроэпизоды в корпусе повседневной русской речи «Один речевой день»: Принципы аннотирования и результаты статистической обработки [в:] *Труды Международной конференции «Корпусная лингвистика – 2013»*, Санкт-Петербург: Филологический факультет СПбГУ, с. 449–456.

Шерстинова Т. Ю., Рыко А. И., Степанова С. Б., 2009, Система аннотирования в звуковом корпусе русского языка «Один речевой день» [в:] *Формальные методы анализа речи: Материалы XXXVIII Международной филологической конференции*, Санкт-Петербург: Факультет филологии и искусств СПбГУ, с. 66–75.

Щерба Л. В., 1974, О частях речи в русском языке [в:] Л. В. Щерба, *Языковая система и речевая деятельность*, Ленинград: Наука, с. 77–100.

Asinovsky A. S., Bogdanova N. V., Rusakova M. V., Ryko A. I., Stepanova S. B., Sherstinova T. Yu., 2009, The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation [в:] V. Matoušek, P. Mautner (eds.), *Text, Speech and Dialogue. TSD 2009*, (LNCS/LNAI 5729), Berlin–Heidelberg: Springer-Verlag, с. 250–257.

British National Corpus, 2007, http://www.natcorp.ox.ac.uk (дата обращения: 14.10.2015).

Hellwig B., Van Uytvanck D., Hulsbosch M. et al., 2014, ELAN – Linguistic Annotator. Version 4.7.0, http://www.mpi.nl/corpus/html/elan/ (дата обращения: 14.10.2015).

Stepanova S., 2013, Speech Rate as Reflection of Speaker's Social Characteristics [в:] N. Telemann, P. Kosta (eds.), *Approaches to Slavic Interaction*, (Dialogue Studies, Vol. 20), Amsterdam–Philadelphia: John Benjamins Publishing Company, с. 117–129.