О корпусе текстов живой речи: новые поступления и первые результаты исследования[1]

The corpus of spoken Russian: new receipts and the first results of research

Богданова Н. В. (nvbogdanova_2005@mail.ru)

Филологический факультет Санкт-Петербургского государственного университета; Санкт-Петербург, Россия

В докладе представлены новые поступления в сбалансированную часть Звукового корпуса русского языка — массива текстов живой монологической речи, объединенных едиными лингвистическими, социолингвистическими и психолингвистическими параметрами. Описываются новые блоки такого корпуса и первые результаты исследования его материала.

Корпус текстов живой монологической речи (текстотека бытовых монологов) представляет собой один из блоков Звукового корпуса русского языка (ЗКРЯ), работа над которым ведется на филологическом факультете СПбГУ. Отличительной особенностью этого блока является достаточно строго сбалансированный характер его формирования. Принцип, положенный в основу организации корпуса, условно можно назвать «*принципом ковчега*» («каждой твари по паре»): балансировке подвергается и состав информантов (с точки зрения их социальных и психологических характеристик), и та лингвистическая программа, по которой осуществляется запись их речи (чтение и пересказ двух текстов, описание двух изображений и свободный рассказ — 7 текстов от каждого информанта) (подробнее см.: *Богданова и др.* 2008).

Работа над этой частью корпуса продолжается в настоящее время по двум направлениям: пополнение его новыми текстами и анализ существующего материала. Целью настоящего сообщения является как раз представление новых структурных частей корпуса и первые конкретные результаты его анализа.

На сегодняшний день структура корпуса выглядит следующим образом — см. таблицу 1:

По сравнению с предыдущим составом, текстотека пополнилась следующим образом:

- два блока интерферированной русской речи иностранцев — американцев и китайцев. Подобный материал дает возможность сравнивать специфические черты спонтанной русской речи говорящих на родном и неродном языке, выявлять универсальные черты спонтанности (в рамках различных коммуникативных сценариев)

Таблица 1. Сбалансированный материал Звукового корпуса русского языка

MED	Речь медицинских работников	32 диктора, 210 текстов	6 часов звучания
JUR	Речь юристов	40 дикторов, 322 текста	16 часов звучания
RKI	Речь преподавателей РКИ	20 дикторов, 70 текстов	3,5 часа звучания
STUD	Речь студентов	5 разных блоков	7 часов звучания
COMP	Речь «компьютерщиков»	12 дикторов, 32 текста	72 мин звучания
RIA	Интерферированная русская речь американцев	68 дикторов, 204 текста	10 часов звучания
RIK	Интерферированная русская речь китайцев	2 диктора, 4 текста	10 минут звучания

[1] Исследование выполнено при поддержке гранта РФФИ «Изучение зависимости речевых характеристик от условий коммуникации (корпусное исследование на материале повседневной русской речи)» (проект 10-06-00300).

и те, что обусловлены интерференционными процессами, возникающими при контакте двух конкретных языков. Насколько можно судить, межъязыковая интерференция на уровне спонтанного речепроизводства вообще редко становилась предметом лингвистического анализа;

- блок профессионально окрашенной бытовой речи «компьютерщиков» (программисты, системные администраторы, преподаватели информатики, студенты, обучающиеся по соответствующим специальностям). Привлечение подобного материала расширяет возможности описания одного из типов внутриязыковой интерференции — между литературной и профессиональной речью носителей языка. В рамках настоящего корпуса это дает возможность сравнивать специфику построения бытового спонтанного монолога в рамках того или иного коммуникативного сценария информантами из разных профессиональных групп: юристами, медиками, преподавателями русского языка как иностранного (РКИ) и «компьютерщиками»;
- небольшой самостоятельный блок речи студентов — точнее, одного информанта, записавшего в течение трех лет полную лингвистическую программу четырехкратно, с месячным интервалом между записями одного и того же коммуникативного сценария (всего 28 текстов). Такой подход позволяет выявить устойчивые черты речевого портрета говорящего, не зависящие от условий протекания конкретного речевого акта, а также степень влияния этих условий на речевую продукцию человека.

Еще одна возможность расширения данной части звукового корпуса и в целом развития исследований в этом направлении видится в пересечении материала двух блоков: данного сбалансированного и корпуса повседневной речи носителей языка «Один речевой день» (ОРД), построенного, в отличие от первого, по «принципу невода» (см. о нем подробнее: *Асиновский и др*. 2008, 2009, а также статью того же авторского коллектива в настоящем сборнике).

Так, возможность сравнить речь человека в разных условиях записи предоставляет эксперимент, осуществляемый в настоящее время: информант-студент, речь которого записывалась многократно в течение трех лет, по одной и той же лингвистической программе (см. выше), записал еще и свой речевой день. Даже с учетом того факта, что он сам прожил этот день «с диктофоном на шее», т. е. знал о проводящейся записи, условия этой записи намного более естественны по сравнению с речью в микрофон. Наши наблюдения показывают, что информант, записывающий свой речевой день, помнит о диктофоне не более первых двух часов, а дальше ведет себя совершенно естественно. В то время как запись с микрофоном и присутствующим экспериментатором ни на секунду не дает информанту расслабиться и забыть о том, что его речь не только слушают, но и фиксируют для дальнейшего анализа. В перспективах этого направления исследования — повторные (как минимум, еще одна, через год) записи речевого дня данного информанта, для выявления всех особенностей его речевого поведения и создания его речевого портрета.

Что касается конкретных результатов исследования материалов этой части ЗКРЯ, то они весьма любопытны и связаны с самыми разными уровнями анализа.

Так, было проведено исследование спонтанной речи информантов из двух профессиональных групп — медиков и юристов, с целью выявления влияния профессии на их бытовую речь. Оказалось, что в обоих случаях это влияние имеет место, но оно принципиально различно. Медиков «выдает» пристрастие к медицинской терминологии, они прибегают к ней тем чаще, чем выше уровень речевой компетенции (УРК) информанта. Более того, информанты с высоким УРК (врачи-преподаватели, профессионально связанные с речью) используют (в рассказе об отдыхе!) достаточно редкие и узко специальные термины (*атеросклероз сосудов головного мозга, сердечная недостаточность, энергозатраты, генотип*), в то время как информанты с низким УРК (медсестры) ограничиваются в своих рассказах терминами, весьма частотными в речи носителей русского языка (*боль, больной, медсестра*). Крайним проявлением влияния профессии на бытовую речь человека можно считать полное переключение говорящего (в рассказе об отдыхе!) на медицинскую тему. Таких случаев в материале выявлено 5 на 30 текстов, при этом ни одного — в монологах информантов с низким УРК. Порой информанты сами осознают эту смену тематики и даже предупреждают об этом собеседника (экспериментатора):

*если бы не Макдоналдс я бы там наверное не выжила потому что японцы едят совершенно изумительную пищу // **э-э опять-таки медицина** // **ну ха-ха вам придется на это сделать сноску** // э-э они где-то лет двадцать назад стали вымирать от инсультов потому что у них произошла американизация и появилась американская пища сеть Макдоналдс все вот эти Кэрролс и так далее // японский организм оказался к этому непривычным и они стали / э-э стали болеть атеросклерозом причем вот именно японская нация стала болеть атеросклерозом сосудов головного мозга // поэтому количество инсультов превзошло превзошло все возможные ожидания они увеличились на порядок по отношению к первоначальному / после этого они вернулись к традиционной пище // традиционную пищу есть могут только японцы* [Инф. 20, жен., высокий УРК].

В целом полученные цифры невелики — около 1 % медицинских терминов на весь лексиче-

ский объем свободных рассказов 30 информантов (4 670 фонетических слов). Однако тема рассказа — о способе проведения свободного времени — вовсе не предполагала использования медицинской терминологии. И взятые для сравнения свободные рассказы информантов-юристов на ту же тему выявили на порядок меньше медицинских терминов (0,07 % от общего числа фонетических слов в текстах), и только широко распространенных — *организм*, *самочувствие*, *реанимация*, *хондроз*, *здоровье*, *инстинкт*, *ангина*. Ни одного случая полного переключения на медицинскую тему в монологах юристов, разумеется, не выявлено.

Влияние профессиональной речи юристов на их бытовую речь оказалось совсем иного рода. Здесь скорее можно говорить о «стилевой разноголосице», о проникновении элементов официально-делового стиля, столь характерного для юридического языка, в бытовые рассказы об отдыхе, ср.:

- *отдых в выходные связан / непосредственно с нахождением дома*;
- *основополагающим моментом моих выходных это должно быть конечно хорошая погода*;
- *их было достаточно много лиц / которые говорили / что нам холодно*;
- *может где-то даже ∫ процентов семьдесят / по моим ∫ э э подсчетам / и наблюдениям / таким образом / отдыхают от ∫ рабочих будней насколько мне известно / больше половины / процентов от всего населения <…> таким образом отдыхают от рабочих будней*;
- *сам процесс отдыха основная составляющая которого / заключается в том что кататься на лыжах*;
- *отдыхаем как совместно с ребёнком / так и не совместно с ребёнком.*

Смешение стилей зачастую рождало в монологах юристов даже комический эффект:

- *лёгкий просмотр телевизора с приёмом завтрака;*
- *долгое / времяпрепровождение вечером с друзьями / в баре;*
- *неприятные ощущения в виде ожогов;*
- *горнолыжный курорт / с очень приятной компанией / с приятным местонахождением;*
- *применяет на себе там / эти косметические средства;*
- *иногда даже / ну употребляется некое количество алкоголя;*
- *в индивидуальном порядке в номере занимаюсь иногда / лечебной такой гимнастикой*.

Два перцептивных эксперимента (на орфографических расшифровках материала и на звучащей речи) показали, что носители русского языка (эксперты-филологи, студенты и преподаватели, 20 чел.) довольно хорошо чувствуют специфику бытовой речи юристов и с высокой степенью вероятности (78–89 % при чтении расшифрованных отрывков и 73–69 % при аудировании) относят предложенные им фрагменты монологов именно к речи юристов (подробнее об этом см.: *Иванова* 2008, 2010). Правда, столь же высок процент отнесения этих фрагментов к профессии менеджера, что позволяет говорить о заметной специфике бытовой речи носителей языка, связанных с деловой сферой жизни.

В ходе анализа материала всего корпуса спонтанных монологов удалось подтвердить справедливость введения в научный обиход такого социологического параметра, как *профессиональное/непрофессиональное отношение говорящего к языку/речи*, и связанного с ним признака *уровня речевой компетенции* говорящего. Анализ речи медиков показал, что даже словарный запас информантов с разным УРК весьма существенно различается: на высоком уровне он составляет около 4 тыс. лексем, на среднем — 3 тыс., на низком — 2 тыс. Другими маркерами УРК говорящего (и все они получили достаточно весомое подтверждение в ходе анализа материала разных блоков корпуса) оказались следующие:

- *степень членимости* текста на единицы, соотносимые с предложением: относительная легкость членения маркирует высокий УРК; наиболее эффективно данный признак диагностирует УРК говорящего на трудных коммуникативных сценариях — в пересказах и описаниях (см. подробнее: *Бродт* 2007);
- *средняя длина «предложений» в словах*: высокому УРК соответствуют длинные «предложения», низкому — короткие (см. там же);
- *употребление вставных конструкций* классического типа (см. *Богданова* 2010б);
- *разнообразие синтаксических конструкций*: диагностирующим УРК является количество сложных «предложений», доля безличных и инфинитивных конструкций и ряд других;
- *разнообразие заполнения синтаксических позиций*, в частности функции инфинитива и употребление причастий и деепричастий (подробнее о маркерах УРК говорящего см.: *Богданова* 2010а).

На обширном материале удалось показать, что синтаксическое и интонационное членение спонтанных текстов разных типов соответствуют друг другу только на 54 % (см. подробнее: *Степихов* 2005).

Анализ *чтения* (корпус речи студентов), несмотря на высокую степень его лингвистической мотивированности первичным текстом, позволил с уверенностью отнести этот вид монолога к разновидностям спонтанной речи. При неподготовленном чтении в речи информантов возникают паузы хезитации, самоперебивы, повторы и различные ошибки — т. е. черты, свойственные любой устной речи:

- *Берестов выехал прогуляться верхом / на всякой случай взяв с собою пары* **<…>** *т***<…>***ри борзых;*
- *огражденный своим чином токмо <…> токмо от побоев; чьи это <…> чьи это дрожки?*
- *и рассказал все / что случи<…> рассказал все / что случилось*; *но он наехал на Берестова вовсе неожи<…> вовсе неожиданно.*

Более того, оказалось, что любое чтение, как неподготовленное, так и подготовленное (разумеется, в рамках лингвистического эксперимента), обладает полным набором черт спонтанности, которые зависят от индивидуальных характеристик говорящего в той же мере, как и любая другая речевая продукция (см. об этом подробнее: *Сапунова* 2009).

На материале монологов-*описаний* (корпус речи студентов) выяснилось, в частности, что в понимании говорящего задача описания изображения существенно расширяется за счет того, что так или иначе связано с изображением: его автор, жанр, искусство в целом, даже собственный опыт человека (подробнее о монологах-описаниях см.: *Филиппова* 2010). В результате отчетливо выделились три класса сценариев:

1) собственно описание:

А) называние, перечисление объектов изображения или событий виртуального мира:
- *летний пейзаж идет тропинка цветочки* [несюж., девушка-нефилолог];

Б) установление отношений объектов или событий с внетекстовой реальностью; суждения, домыслы, догадки говорящего:
- *на шестой* [картинке] *какой-то дядька пожарник судя по всему снимает ее с лестницей* [сюж., юноша-филолог];

2) метакоммуникация — речь информанта о собственной речи («текст о тексте») или о самой ситуации общения:
- *я много не умею говорить / могу красиво молчать* [несюж., юноша-нефилолог];
- *что бы еще вам такого описать* [несюж., юноша-нефилолог].

3) комментирование — речь информанта о собственном опыте, об ассоциациях и т. п.:
- *но он / у нас пейзажист был* [о Шишкине] *так что в общем что с него взять* [несюж., девушка-нефилолог];
- *вообще мне картины Шишкина очень нравятся / еще помню спор был сколько там на картине / сколько там мишек было* [несюж., юноша-филолог].

Пересказы информантов-юристов также позволили выявить различные коммуникативные стратегии, используемые в репродуцированных текстах говорящими с разными социальными и психологическими характеристиками (подробнее о монологах-пересказах см.: *Куканова* 2009).

Я-наблюдатель (наиболее распространенная стратегия — 63 % порожденных текстов, более свойственна для говорящих с высоким УРК, интровертов) — стремление информанта изложить более или менее точно своими словами содержание прочитанного отрывка, передав в сюжетном тексте событийную линию, а в несюжетном — свойства и состояния объектов, которые описываются в первичном тексте:
- *был тёплый июльский день // солнце поднималось ∫ на восходе // солнце было не ∫ не яркое / не огненно-рыжее / оно было светлое и лучезарное // лучи его поднимаясь над землей / пронизывали* (…) *светлые / белые облака // облака <вздох> стоящие над землей* (…) *вверху на небе / лежали белыми белыми яркими ∫ снопами* [Текст 2. Инф. 12].

Я-читатель (32 % проанализированных текстов, самая простая стратегия, в наибольшей степени представлена в монологах говорящих с низким УРК) — стремление информанта передать тематическое содержание первичного текста с выделением главной (доминирующей) темы и ее компонентов:
- *ну / начинается с того что* […] *с описания ∫ (э-э) женщины и пса // входят они / оказываются в какой-то комнате / и реакция пса / вот на ту ситуацию которую он видит // там / какая-то лестница темная* [Текст 1. Инф. 16].

Я-исследователь — (наименее востребованный способ организации текста-пересказа — около 5 %, характерен для мужчин младшей возрастной группы с высоким УРК, экстравертов) попытка информанта проанализировать первичный текст:
- *в общем-то ∫ в данном тексте говорится о ∫ (а-а) ∫ как мы понимаем / о женщине / которая привела собаку / в некую комнату // при этом ∫ (а-а) параллельно видя ∫ (э-э) ситуацию глазами ∫ (а-а) стороннего наблюдателя и глазами собаки* [Текст 1. Инф. 14].

На материале репродуцированных текстов удалось выявить также способы лексического наполнения вторичного текста в сравнении с первичным. *Эндоединицы* полностью повторяют единицы предтекста, а *эзоединицы* — это различные новые единицы, появившиеся только в пересказе, но мотивированные исходным текстом тем или иным способом. Среди последних выявлены, например грамматические и семантические трансформеры (*пощелкала пальцами — пощелкав пальцами, сладкий — сладковатый*), транспозиты (*кричал — крики, солнце — солнечный*), синонимы (*синева — лазурь, земледелец — хлебопашец*), антонимы (*сырость — сухость*), гипонимы/гиперонимы (*пес — животное, комната — помещение*), конверсивы (*она <…> наполнила комнату запахом — комната наполнилась*

запахом), свернутые/развернутые номинации (*личность мужского пола — мужчина, облака с белыми краями — края [облаков] были белые*), ассоциативные замены разного типа (*касторка — валерьянка, небосклону — небоскребу, осколки — обломки, стекло + дверь — окно*), вводящие элементы текста (*повествуется*, *описывается, я прочитал отрывок из Тургенева*), слова, выражающие эмоции (*бедолага пес*) или модальную оценку (*скажем так пожалуй, во-первых, во-вторых*), выполняющие метакоммуникативную (*да сложновато так, насколько я помню, да неправильно*) или дискурсивную функцию (*всё* — в конце монолога), и т. п.

Корпусный характер организации материала позволяет с помощью специальных программ создавать *конкордансы* (алфавитно-частотные словники использованных информантами лексических единиц) разного типа: как общие по разным типам текстов, так и по отдельным группам информантов. По этим словникам можно видеть высокую употребительность таких специфических для спонтанной речи не вполне речевых элементов звуковой цепи, как хезитативы *э-э* или *а-а*, или частиц *вот* и *ну*, также зачастую выполняющих в спонтанной речи не служебную, а чисто хезитационную функцию. Знаменательные части речи уступают таким элементам по частоте употребления порой в десятки раз. Кроме того, появляется возможность сравнивать частотность в устной речи различных грамматических форм одного и того же слова и видеть, например, преобладание начальной формы существительного над косвенными и, наоборот, существенное преобладание личных форм глагола над исходной (инфинитивом). Можно сравнивать между собой и лексикон разных групп носителей языка (организованных по гендерному, возрастному, профессиональному, психологическому и т. п. признакам) в сходных коммуникативных условиях и решать еще множество других исследовательских задач.

Так, в монологах-описаниях среди самых частых слов в словниках практически всех групп информантов (студенты — филологи и нефилологи, юноши и девушки) — отрицательная частица *не*, в какой-то степени свидетельствующая о своеобразной «борьбе» говорящего с трудным коммуникативным сценарием. Значительная часть контекстов с этим *не* — конструкция *не знаю*. А в описаниях несюжетного изображения (самый трудный сценарий) в «верхушку» частотников попало и слово *наверное*, также отражающее поиск говорящими нужного слова, наряду с крайне частотными хезитационными словечками *ну*, *вот*, *это*, *там* и под. Единственным полноценным словом в этой части словника оказалось наречие *очень*, использованное девушками-нефилологами.

Началась и публикация материалов сбалансированной части ЗКРЯ (см. *Русская спонтанная речь* 2008, 2010а,б), что делает их доступными широкому кругу исследователей самого разного ранга.

Литература

1. *Асиновский А. С., Богданова Н. В., Русакова М. В., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008 С. 488–494.
2. *Асиновский А. С., Богданова Н. В., Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 8 (15). По материалам ежегодной международной конференции «Диалог» (2009) / Гл. ред. А. Е. Кибрик. М., 2009. С. 38–44.
3. *Богданова Н. В.* Уровень речевой компетенции как реальная социальная характеристика говорящего, определяющая его речь // Материалы XXXVIII международной филологической конференции. Выпуск 22. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 16–20 марта 2009 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2010а. С. 29–40.
4. *Богданова Н. В.* Вставные конструкции в звучащем спонтанном монологе (к проблеме построения грамматики русской речи) // Вопросы культуры речи. М., 2010б (в печати).
5. *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. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 57–61.
6. *Бродт И. С.* Спонтанный монолог в лингвистическом и социолингвистическом аспектах (на материале текстов разного типа). Дис. … канд. филол. наук. СПб., 2007 (машинопись).
7. *Иванова О. А.* К характеристике внутриязыкового контакта между литературной и профессиональной речью носителя русского языка // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 25–35.
8. *Иванова О. А.* Стилевая «разноголосица» в бытовой спонтанной русской речи // Вестник Санкт-Петербургского университета. Филология. Востоковедение. Журналистика. Серия 9. СПб., 2010 (в печати).
9. *Куканова В. В.* Лингвистический анализ репродуцированных текстов (на материале звукового корпуса русской речи юристов). Дис. … канд. филол. наук. СПб., 2009 (машинопись).
10. *Русская спонтанная речь.* Свободные монологи-рассказы на заданную тему. Тексты. Лексические материалы / Сост. В. В. Куканова / Отв. ред. и автор предисловия Н. В. Богданова. СПб., 2008.
11. *Русская спонтанная речь.* Монологи-репродуктивы. Тексты. Лексические материалы / Сост. В. В. Куканова / Отв. ред. и автор предисловия Н. В. Богданова. СПб., 2010а (в печати).
12. *Русская спонтанная речь.* Монологи-описания. Тексты. Лексические материалы / Сост. В. В. Куканова / Отв. ред. и автор предисловия Н. В. Богданова. СПб., 2010б (в печати)/
13. *Сапунова Е. М.* Неподготовленное чтение как вид речевой деятельности и тип устного спонтанного монолога (на материале русского языка) Дис. … канд. филол. наук. СПб., 2009 (машинопись).
14. *Степихов А. А.* Соотношение синтаксического и интонационного членения в спонтанном монологе. Дис. … канд. филол. наук. СПб., 2005 (машинопись).
15. *Филиппова Н. С.* Принципы построения устного описательного дискурса (на материале русской спонтанной речи). Дис. … канд. филол. наук. СПб., 2010 (машинопись).