О «КОРПУСЕ» ТЕКСТОВ ЖИВОЙ РЕЧИ: ПРИНЦИПЫ ФОРМИРОВАНИЯ И ВОЗМОЖНОСТИ ОПИСАНИЯ

THE CORPUS OF SPOKEN RUSSIAN: DESIGN PRINCIPLES AND APPROACHES TO DATA ANALYSIS

Богданова Н.В. *(nvbogdanova_2005@mail.ru),* ***Бродт И.С.*** *(brodt_05@mail.ru),* ***Куканова В.В.*** *(vika.kukanova@gmail.com),* ***Павлова О.В.*** *(olgapavlovaspb@mail.ru),* ***Сапунова Е.М.*** *(kaverita@yandex.ru),* ***Филиппова Н.С.*** *(ninaphilippova@gmail.com)*
Санкт-Петербургский государственный университет

В докладе рассматриваются принципы формирования своеобразного звукового корпуса – массива текстов живой русской речи, объединенных едиными лингвистическими и социолингвистическими параметрами. Описываются готовые блоки такого корпуса, возможности его многоуровневого описания и перспективы расширения.

Одним из актуальных и активно развивающихся направлений современной лингвистики является сбор и систематизация живого речевого материала[1]. Этим занимается *полевая лингвистика*, под которой понимают «лингвистическую дисциплину, разрабатывающую и практикующую методы получения информации о неизвестном исследователю языке на основании работы с его носителями» (Кибрик 2007). В настоящем исследовании методы полевой лингвистики используются для построения и описания своеобразного «корпуса» текстов живой русской речи, о которой, как показывают первые исследования на этом материале, лингвистике известно примерно так же мало, как о каком-нибудь действительно неизвестном языке, и любая попытка описать эту речь с применением тех методов и того лингвистического инструментария, который традиционно используется для анализа письменно-литературного языка, наталкивается на сопротивление самого материала, что вынуждает исследователей ставить и решать массу абсолютно новых задач, начиная с определения самóго метаязыка такого лингвистического описания (см. *Полевая лингвистическая практика 2008*). Иными словами, к спонтанной речи – с большей или меньшей степенью уверенности – мы можем относиться, как к новому объекту изучения, и применять к нему устоявшиеся методы полевой лингвистики. А. Кибрик противопоставляет полевую лингвистику «кабинетной», на том основании, что источником данных для последней «являются либо языковая интуиция самого исследователя, являющегося носителем изучаемого языка или, по крайней мере, хорошо им владеющего, либо обширный корпус текстов на изучаемом языке, о котором опять же известно достаточно много для того, чтобы изучать его без обращения к суждениям его носителей» (*Кибрик 2007*). В случае с живой речью, с одной стороны, исследователи сами являются ее носителями, и даже «хорошо ею владеющими», но, с другой, не могут изучать ее без обращения к другим носителям. Именно эти неискушенные носители языка – информанты – являются «посредниками между исследователем и языком», а задача исследователя – «эффективно воздействовать на языковую деятельность информанта», чтобы получить от него тот или иной речевой продукт. Для получения такого продукта полевая лингвистика использует «активный метод целенаправленного интервьюирования по определенной программе», в ходе осуществления которого языковая деятельность информанта протекает максимально естественно и спонтанно.

Именно такая программа легла в основу формирования звукового «корпуса» спонтанных текстов на русском языке, создаваемого нашим научным коллективом. Слово «корпус» употреблено в данном случае до некоторой степени условно, речь идет скорее о своеобразной текстотеке, о массиве живых текстов на русском языке, у которых есть ряд общих и ряд различных черт и которые могут быть, с одной стороны, материалом для самых разных исследований (от фонетики до функциональной стилистики, в том числе в социо- и психолингвистическом аспектах), а с другой – частью того самого корпуса живой русской речи, создание которого и является целью современной полевой лингвистики.

[1] См. об этом: *Полевая лингвистическая практика 2007.* В качестве примера таких баз данных можно назвать, кроме того, проект «Отчеты детей об их сновидениях» (руководители А. А. Кибрик и В. И. Подлесская)», а также устную часть Национального корпуса русского языка.

С точки зрения функциональной принадлежности можно говорить о монологах различного характера. Чаще всего исследователи обращают свой взор на устную публичную речь – общественно-политическую, деловую или научную. В противоположность этому нас интересует только *бытовой спонтанный монолог*, характеризующийся неподготовленностью, непринужденностью, неофициальностью и необязательным участием говорящего в акте коммуникации (роль второго участника такого акта сводится к минимуму, хотя, безусловно, никогда не исчезает полностью).

Непременное требование спонтанности, предъявляемое к материалу фиксации и исследования, не отменяет того факта, что степень такой спонтанности может быть различной и зависит от многих факторов.

Главным из таких факторов следует назвать степень *лингвистической мотивированности* речевого произведения, то есть обусловленности (тематической и лингвистической) монолога как продукта речевой деятельности человека (вторичного текста) характеристиками некоторого исходного, первичного, ставшего стимулом для появления вторичного. Разнообразие таких стимулов велико: от вопроса, ответом на который становится развернутый спонтанный монолог (минимальная степень мотивированности и максимальная степень спонтанности), до другого текста (предтекста), предназначенного для прочтения (максимальная степень мотивированности и минимальная степень спонтанности) или пересказа (более низкая, чем при прочтении, степень лингвистической мотивированности и, соответственно, более высокая степень спонтанности). Промежуточное положение на этой шкале занимает описание зрительного ряда (изображения), обладающее средней степенью мотивированности и такой же средней степенью спонтанности.

Лингвистическая мотивированность и спонтанность речевого произведения (в нашем случае – бытового монологического текста) находятся в отношениях обратно пропорциональной зависимости: увеличение степени мотивированности ведет к снижению спонтанности, и наоборот[2].

Дополнительными характеристиками исходного стимула, способными повлиять на свойства спонтанного монолога, стали в нашем исследовании сюжетность/несюжетность предтекста или изображения или степень знакомства говорящего с темой свободного монолога, заданного вопросом (исходным стимулом). Эти дополнительные характеристики не меняют степени лингвистической мотивированности и, соответственно, спонтанности вторичного текста, но все же оказывают влияние на выбор говорящим речевых средств и в целом на лингвистическую природу вторичного текста. Можно предположить, что в данном случае решающими являются характеристики уже не (или не только) первичного текста, но и самого говорящего – уровень его речевой компетенции (УРК) или психологический тип его личности.

Наш корпус, таким образом, составляют спонтанные бытовые монологи разной степени лингвистической мотивированности и спонтанности, записанные от носителей русского языка с различными социальными характеристиками. Последние представлены главным образом такими показателями:

- пол,
- возраст,
- профессиональная принадлежность,
- профессиональное или непрофессиональное отношение к речи,
- уровень речевой компетенции.

Думается, что требуют комментариев два последних признака.

Профессиональное или непрофессиональное отношение говорящего к речи устанавливается через определение роли речи (языка) в его жизни. Здесь возможно несколько вариантов:

1) речь для человека – только *средство коммуникации* (большинство носителей языка – не школьники, не филологи, не преподаватели, не актеры, не лекторы и т. п.);

2) речь – *средство коммуникации и объект изучения* (до некоторой степени школьники, а также студенты-филологи и «кабинетные» ученые-филологи);

3) речь – *средство коммуникации* и *орудие труда* (преподаватели-нефилологи, актеры, лекторы, публичные и общественные деятели, политики и т. п.)[3];

4) речь – *средство коммуникации, объект изучения* и *орудие труда* (преподаватели-филологи; особое место среди них, как представляется, занимают преподаватели русского языка как иностранного).

Уровень речевой компетенции говорящего определяется целой совокупностью его социальных характеристик, среди которых ведущее место занимают уровень образования, профессиональное или непрофессиональное отношение к речи, а также степень социальной активности личности. Этот набор признаков, определяющих УРК, был в свое время установлен экспериментальным путем – через экспертную оценку большого массива звучащих

[2] См. подробнее об этой типологии Богданова 2004; Полевая лингвистическая практика 2008.

[3] Число таких носителей языка в современном мире неуклонно увеличивается, что ставит перед специалистами отдельную задачу обучения устной публичной речи. Этой цели может служить, в частности, и создание различных корпусов (текстотек) живой речи, записанных, в частности, от носителей языка с высоким УРК.

текстов и выявление корреляции этих оценок с реальными социальными характеристиками говорящих[4]. В нашем случае мы просто опирались на эти признаки и на их основе определяли УРК наших информантов, хотя проведение экспертной оценки и уточнение проведенного разделения на группы вполне возможно.

Общую характеристику нашего материала удобно дать, отталкиваясь от определения *национального корпуса*, принятого в современной корпусной лингвистике, – это «информационно-справочная система, основанная на собрании текстов в электронной форме. Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всем многообразии жанров, стилей, территориальных и социальных вариантов и т. п. <…> Национальный корпус создается лингвистами (специалистами по так называемой корпусной лингвистике, быстро развивающейся современной области языкознания) для научных исследований и обучения языку. <…> Национальный корпус имеет две важные особенности. Во-первых, он характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленных в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию). Разметка – главная характеристика корпуса; она отличает корпус от простых коллекций (или “библиотек”) текстов. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса»[5]. Приложим конкретные характеристики национального корпуса к нашему материалу.

1. *«Информационно-справочная система, основанная на собрании текстов в электронной форме»*. Наш материал представляет собой некоторую основу для настоящей *«информационно-справочной системы»*, – это ***собрание текстов*** в виде магнитных записей и их орфографических расшифровок. Часть материала существует уже и в электронном виде, возможность перевода в эту форму остальных текстов, разумеется, существует.

2. Корпус представляет *«…язык во всем многообразии жанров, стилей, территориальных и социальных вариантов и т. п.…»*. Выше уже была представлена наша ***типология текстов***, не претендующая, безусловно, на всеохватность, но соответствующая тем теоретическим установкам, на которых строится анализ собираемого нами материала.

3. Корпус характеризуется *«представительностью, или сбалансированным составом текстов»*. В рамках выдвинутой гипотезы о существовании корреляции между лингвистическими характеристиками спонтанного монолога и его типом, с одной стороны (собственно лингвистический аспект исследования), а также между этими характеристиками и социальными и психологическими признаками говорящих, с другой (психо- и социолингвистический аспекты исследования), вполне можно, думается, говорить о ***представительности*** и ***сбалансированности***.

4. Корпус *«содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию)»*. В нашем случае такой дополнительной информацией (и одновременно разметкой) является прежде всего *синтаксическое пунктирование текстов*, полученное для части материала в ходе специальных экспериментов с привлечением экспертов-филологов. Подобное пунктирование дает в руки исследователю некую единицу описания, соотносимую с традиционным предложением и позволяющую осуществить дальнейший синтаксический анализ спонтанных монологов во всей его полноте[6].

Другим вариантом синтаксической разметки звучащего материала стало выделение в спонтанном тексте *стуктурно-синтаксических единств* (ССЕ), под которыми понимаются связанные комплексы – «предикативные центры (полнозначные слова в функции главных членов предложения и нечленимые слова-предложения), сами по себе или с зависимыми словоформами и полупредикативными конструкциями» (*Филиппова 2006*).

Другим типом разметки на нашем материале является отражение в расшифровках того, что в определении корпуса обозначено как *«изменение качества речи, паузация и разнообразные паралингвистические явления (например, смех) в устной речи»*. Все это (а также повторы, обрывы речи, самоперебивы и самокоррекция, паузы хезитации – как неотъемлемые признаки любого спонтанного монолога) присутствует в зафиксированном материале, а в некоторых случаях даже стало объектом специального рассмотрения.

Наличие в нашем материале подобной разметки уже позволяет говорить о нем, как о некоем подобии корпуса, поскольку именно *«она отличает корпус от простых коллекций (или “библиотек”) текстов»*.

5. Наш материал представляет собой своеобразное объединение того, что в определении корпуса называется *«демографической частью»* (спонтанная речь повседневного общения) и *«контекстно-ориентированной» устной речью*, – см. предложенную выше типологию составляющих его спонтанных монологов.

6. Наш материал – это, безусловно, *«корпус общего типа»*, т. к. он содержит разные речевые жанры.

[4] См. в списке использованной литературы серию отчетов ЛЭФ им. Л. В. Щербы за 1985-90 гг.

[5] Что такое Корпус? // Официальный сайт Национального корпуса русского языка. http://www.ruscorpora.ru/corpora-intro.html.

[6] См. об этом подробнее *Богданова 2006, Бродт 2007*.

В целом возможные *аспекты использования* собранного материала можно представить следующим образом:

1) собственно *лингвистические исследования*:

• специфика устной спонтанной речи на всех уровнях;

• пересмотр нормативных требований к построению живого (в первую очередь устного) монологического текста;

• создание лексикографического описания бытовой спонтанной звучащей речи;

• описание дистрибуции тех или иных грамматических классов слов или их форм в устной монологической речи разных социальных групп;

2) *лингводидактика*:

• обучение русскому языку нерусских; собрание дает богатый материал для учебного аудирования и вообще знакомства с живой речью в иностранной аудитории;

• изучение грамматики речи в русской филологической аудитории;

3) материал для *психо- и социолингвистических исследований*;

4) материал для исследований в области *коллоквиалистики*;

5) *прикладная лингвистика*:

• решение задач обработки естественного языка/речи;

• решение задач интегрального моделирования звуковой формы.

Основная единица описания в собрании – *спонтанный монологический текст* в звучащем и расшифрованном виде.

Состав информантов – носителей русского языка, от которых записан весь наш материал, – будучи весьма разнородным по социальным и психологическим характеристикам говорящих (это было одним из непременных условий всех проведенных экспериментов), является, тем не менее, строго однородным в территориальном отношении: все информанты являются коренными петербуржцами, т. е. носителями петербургского произносительного варианта современного русского литературного языка.

Объем и состав «корпуса»:

1) спонтанные устные монологи разного типа (здесь и далее речь идет о предложенной выше типологии), записанные от 30 информантов, женщин-медиков одной возрастной группы, но с разным УРК (210 текстов, около 9 час. звучания) (*Бродт 2007*);

2) спонтанные устные монологи разного типа, записанные от 6 информантов, преподавателей русского языка как иностранного (наиболее высокий УРК), разного пола и возраста (42 текста, около 2 час. звучания) (*Павлова 2007*);

3) спонтанные устные монологи разного типа, записанные от 43 информантов, мужчин-юристов разного возраста и разного УРК (301 текст, около 12 час. звучания) (*Куканова 2007; 2008*);

4) спонтанные устные монологи – описания сюжетного и несюжетного изображения, – записанные от 20 информантов одного возраста (19 22 года) и приблизительно одного УРК (студенты – филологи и нефилологи) (40 текстов, около 2 час. звучания) (*Филиппова 2006; 2007*);

5) спонтанные устные монологи – свободные рассказы на заданную тему (знакомую и незнакомую), – записанные от 20 информантов одного возраста и приблизительного одного УРК (студенты – филологи и нефилологи) (40 текстов, около 2 час. звучания) (*Конюхова 2006*);

6) неподготовленное (спонтанное) чтение двух текстов (сюжетного и несюжетного), записанное от 12 информантов разного пола, но одной возрастной группы и одного – среднего – УРК (студенты – филологи и нефилологи) (24 текста; 1 час звучания) (*Сапунова 2007)*;

Как видно, общий объем материала, ставшего частью создаваемого массива («корпуса») живых текстов на русском языке, достаточно представителен и разнообразен. Столь же разнообразны и возможности его расширения и практического использования.

Список литературы

1. Богданова Н.В. Типология спонтанных монологов в устной и письменной формах речи // Фонетические чтения. К 100 летию Л. Р. Зиндера. СПб.: 2004. С. 214 217.

2. Богданова Н.В. О единице описания синтаксической структуры устного спонтанного монолога: проблемы, методики, гипотезы // …СЛОВО ОТЗОВЕТСЯ. Памяти А. С. Штерн и Л. В. Сахарного. Пермь: 2006. С. 288 293.

3. Бродт И.С. Спонтанный монолог в лингвистическом и социолингвистическом аспектах (на материале текстов разного типа). Дис. … канд. фил. наук, СПб.: 2007.

4. Исследование отражения в речи некоторых социальных характеристик говорящего. Отчеты ЛЭФ. Л., 1987, 1988, 1990.

5. Кибрик А. Полевая лингвистика // www.krugosvet.ru/articles/77/1007704/1007704a1.htm (2007).

6. Конюхова А.А. Лингвистические особенности свободного монолога на заданную тему // Русская филология. 18. Сборник научных работ молодых филологов. Тарту: 2007. С. 222 230.

7. Куканова В.В. Об одном из способов подбора информантов в ходе полевого исследования // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 45 52.

8. Куканова В.В. Русская спонтанная речь. Методическая разработка по современному русскому языку. Выпуск I. Свободные монологи-рассказы на заданную тему. Тексты. СПб.: 2008 (в печати).

9. Методика получения языкового материала для изучения социальной характеристики говорящего. Отчет ЛЭФ. Л., 1985.

10. Павлова О.В. Влияние возраста говорящего на синтаксические характеристики его речи // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 52 59.

11. Полевая лингвистическая практика. Учебно-методический комплекс сложной структуры. Часть 1. Теоретические основы и методика сбора лингвистических данных для представления их в речевом корпусе русского языка / Ред. Асиновский А. С., Богданова Н. В. СПб., 2007.

12. Полевая лингвистическая практика. Учебно-методический комплекс сложной структуры. Часть 2. Методические указания по обработке, многоуровневой разметке и лингвистическому анализу корпуса звучащих текстов на русском языке / Ред. Асиновский А. С., Богданова Н. В. СПб.: 2008 (в печати).

13. Сапунова Е.М. Неподготовленное чтение как разновидность устного спонтанного монолога // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 76 86.

14. Филиппова Н.С. Опыт анализа синтаксической структуры устного спонтанного монолога-описания // IX Межвузовская научная конференция студентов-филологов. Тезисы. 10 14 апреля 2006 г. Санкт-Петербург. СПб.: 2006. С. 51 52.

15. Филиппова Н.С. Операции отмены как способ организации спонтанной речи (на материале устных спонтанных монологов-описаний) // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12 17 марта 2007 года, СПб.: 2007. С. 86 90.

16. Что такое Корпус? // Официальный сайт Национального корпуса русского языка. http://www.ruscorpora.ru/corpora-intro.html.