**РЕФЕРАТ**

**по проекту № 2.1.3/2023 «Звуковое описание грамматических единиц русского языка: информационно-коммуникативный подход»**

аналитической ведомственной целевой программы “Развитие научного потенциала высшей школы (2009-2010 годы)”

мероприятие: 2 «Проведение фундаментальных исследований в области естественных, технических и гуманитарных наук. Научно-методическое обеспечение развития инфраструктуры вузовской науки»

раздел:2.1 «Проведение фундаментальных исследований в области естественных, технических и гуманитарных наук»

подраздел:2.1.3 «Проведение фундаментальных исследований в области гуманитарных наук»

Руководитель А.С.Асиновский

Отчет 318 с., 5 ч., 87 рис., 54 табл., 38 источников, 9 прил.   
(в томе II – еще 2 приложения)

КЛЮЧЕВЫЕ СЛОВА: звуковой корпус русского языка, грамматика речи, многоуровневая лингвистическая разметка, СПонтанная речь, ФОНЕТИКА, повседневная коммуникация, информационные и речевые технологии

Проект «Звуковое описание грамматических единиц русского языка: информационно-коммуникативный подход» направлен на решение актуальной задачи описания грамматических единиц русской звучащей речи (предложения, словоформы, морфемы) на основе звукового корпуса русского языка повседневного общения с привлечением современных информационных методик.

**Объект исследования или разработки**

Объектом исследования является живая спонтанная речь на русском языке, используемая в повседневном общении – ее грамматика и фонетика. Исследование предполагает анализ ряда социологических факторов (возраст, образование, профессия), психологического типа говорящего и других условий общения в реальных коммуникативных ситуациях.

Объектом разработки является речевой корпус русского языка повседневного общения, составленный из звукозаписей всей речевой деятельности информантов в течение суток («Один речевой день» или ОРД). Корпус интегрируется с информационно-исследовательской средой, состоящей из базы данных, средств лингвистического аннотирования и утилит автоматической обработки данных на фонетическом, лексическом и грамматическом уровнях.

**Цели и задачи проекта**

Устная форма функционирования языка является первичной и определяющей в жизни человека, однако практически все современные грамматики русского языка строились на исследовании письменных источников, без привлечения спонтанной, живой, разговорной речи. Основной целью проекта является описание реальной «грамматики речи» современного русского языка в естественных коммуникативных ситуациях. В частности, предполагается получить описание таких грамматических единиц, как предложение, словосочетание, словоформа, морфема. Осуществление этой цели стало возможно лишь в наши дни при условии привлечения последних достижений в области информационных и речевых технологий.

Предполагается решение следующих задач: 1) создание речевого корпуса качественно нового типа, включающего в себя повседневную устную речь («речевой день») представительной и сбалансированной выборки информантов; 2) осуществление многоуровневой лингвистической разметки речевого материала; 3) реализация процедур автоматического анализа лингвистических и паралингвистических данных, интеграция средств автоматической обработки в единую информационную среду; 4) проведение фундаментальных исследований фонетики и грамматики повседневной речи с использованием средств автоматического анализа; 5) описание отдельных аспектов фонетики и грамматики повседневной речи, представленное в виде коллективной монографии и серии статей.

**Метод или методология проведения работы**

- методы и подходы полевой лингвистики (интервьюирование, анкетирование и др.);  
- методы речевых технологий (комплекс методов звукозаписи, обработки   
и анализа речевых процессов; построение речевых баз данных);

- методы корпусной лингвистики;

- методы лингвистического аннотирования (орфографическая расшифровка, транскрибирование, паузация с дифференциацией длительности физических пауз, синтаксическое пунктирование (экспериментальная расстановка точек), частеречная и синтаксическая разметка и др.;

- статистические методы обработки данных;

- методы социологического анализа, направленные на получение сбалансированной выборки группы информантов;

- методы психологического тестирования;

- информационные методы построения мультимедийных баз данных;

- методика звукозаписи «одного речевого дня» (аудиомониторинг всей речевой деятельности информантов, принимающих участие в реализации проекта, и их коммуникантов в течение одного или нескольких дней).

**Результаты работы**

На втором этапе работы по проекту (3-4 кварталы 2009 г.) в соответствии с календарным планом были решены следующие задачи:

1. Продолжена работа по записи речи новых информантов по методике «одного речевого дня», разработанного коллективом в рамках работы над данным проектом, и их анкетирование. Результаты обработки анкет 40 информантов представлены в базе данных.

2. К настоящему времени записано 320 часов звучания, полученные от 40 информантов (20 мужчин и 20 женщин). Звукозаписи переформатированы, убраны длительные (больше 5 минут) шумовые фрагменты, не содержащие речи. Звукозаписи разрезаны на коммуникативные эпизоды по принципу общих условий коммуникации и качества звукозаписи. В результате было получено 994 файла-эпизода общей продолжительностью 268 часов.

3. Разработана структура речевого корпуса «Один речевой день» (ОРД), состоящая из разных модулей и типов данных: звуковых файлов разных форматов, файлов аннотаций, словарей (частотных, конкордансов), текстовых файлов и др.

4. Подготовлены методики многоуровневого аннотирования данных на лингвистическом и паралингвистическом уровнях. Определены основные типы данных лингвистической разметки, их структурно-системная иерархия. Определены правила членения материала на фонетическом, лексическом, грамматическом уровнях.

5. Осуществлена расшифровка 50 часов звукозаписей для 40 информантов и их многоуровневое аннотирование в профессиональной программе ELAN. В результате расшифровки получены тексты общим объемом в 244075 слов на уровне Frase (реплики говорящих), которые относятся к 125 эпизодам и соответствуют 33,87 часам непрерывной речи. Для 20 информантов-мужчин получено 106109 словоупотреблений на уровне реплик, что соответствует 15,10 часам непрерывной речи, для 20 информантов-жен. Начато сегментирование речевого материала на лексическом уровне. Осуществлена сегментация 14884 слов для 5 информантов (S01, S19, S24, S35, S37), 147 минут звучания.

6. Получена выборочная реальная транскрипция для 1000 словоизменительных морфем для 4 информантов. Морфемная аннотирование включает три уровня – орфографическую запись морфемы, её функциональный тип и реальную фонетическую транскрипцию.

7. Разработан действующий макет специализированной базы данных речевого корпуса в формате MS Access, объединяющей все типы релевантной информации от описания информантов до расшифровок записей. Из среды базы данных возможно прослушивание всех 268 часов звучания - отформатированных ord-файлов.

8. Разработаны алгоритмы автоматической обработки аннотаций, предполагающие использование встроенных средств основных программных продуктов (ELAN, Praat, VBasic/MS Access, STATISTICA) и разработанных творческим коллективом приложений (eCar, eLex и вспомогательные утилиты).

9. Подготовлен частотный словарь словоформ полученных расшифровок. Выполнен первичный анализ данных.

10. Параллельно с созданием корпуса начаты исследования особенностей фонетики и грамматики повседневной речи в зависимости от таких факторов, как социально-психологические особенности говорящего.

11. По мере расшифровки материал поступает в Национальный корпус русского языка, где представлен в подкорпусе «Устный корпус. Из материалов корпуса "Один речевой день", подготовленного группой А.С.Асиновского».

12. Подготовлена демо-версия корпуса, включающая в себя звуковые фрагменты, орфографические расшифровки и файлы аннотаций формата \*.eaf для 10 информантов общей продолжительностью 32 минуты. Демо-версия активно используется в образовательном процессе для студентов филологического факультета СПбГУ и в работе немецких коллег из Университета г. Потсдам. Планируется её публикация в сети Интернет.

13. Основные итоги работы по теме исследования изложены в 26 публикациях.

Объём работ, запланированный на 3-4 кварталы 2009 года (второй этап работы по теме), выполнен полностью.

**Основные конструктивные, технологические и технико-эксплуатационные характеристики**

Речевой материал: 320 часов звучания, 994 отформатированных файла общей продолжительностью 268 часов звучания, формат WAV, 22050 Гц, 16 бит, моно.

Файлы аннотации: 50 часов звукозаписей (33,87 часов непрерывной речи), 125 файлов, формат EAF, 8 уровней сегментации, 244075 словоупотреблений.

Сегментация на фонетические слова: формат Praat-EAF, 1,7 часа, 7 файлов, 9514 фонетических слов.

Сегментация на морфемы: формат Praat-EAF, 5 файлов, 1000 словоизменительных морфем.

База данных: формат MS Access 2003, 14 основных таблиц, возможность озвучивания как всех файлов речевого корпуса, так и изолированно отсегментированных единиц (фраз, слов, морфем).

**Степень внедрения. Итоги внедрения результатов НИР**

По мере обработки материал поступает в Национальный корпус русского языка, где представлен в подкорпусе "Устный корпус. Из материалов корпуса "Один речевой день", подготовленного группой А.С.Асиновского". В НКРЯ уже переданы тексты расшифровок объемом 40000 словоупотреблений.

Звуковой корпус используется как образец на занятиях по спецкурсу «Полевая лингвистика», читаемому на факультете филологии и искусств СПбГУ, в курсах лекций и на практических занятиях в магистратуре по специальности «Полевая лингвистика». На речевом материале корпуса осуществляются исследования в рамках курсовых, дипломных, кандидатских и докторских научных работ.

**Область применения**

Важность и актуальность реализации предлагаемого проекта определяется прежде всего высокой практической значимостью адекватного описания речевого поведения человека в реальном мире и все возрастающей востребованностью современных систем сбора, хранения и анализа речевого материала как для решения фундаментальных научных задач, так и для решения актуальных прикладных задач в области речевых технологий. Среди них:

1) собственно лингвистические исследования:

- описание звуковой формы выявляемых в речи грамматических единиц;

- создание лексикографического описания бытовой спонтанной звучащей речи;

- описание дистрибуции грамматических классов слов и их форм в устной речи разных социальных групп;

2) лингводидактика:

- обучение русскому языку нерусских;

- изучение грамматики речи в русской филологической аудитории;

3) прикладная лингвистика, речевые технологии:

- решение задач обработки естественного языка/речи;

- решение задач интегрального моделирования звуковой формы,

- синтез и распознавание речи.

**Экономическая эффективность или значимость работы**

Экономическая эффективность работы высокая, несмотря на трудоемкость отдельных видов работ – лингвистического аннотирования и сегментации – которые на данный момент принципиально не могут быть автоматизированы. Экономическая значимость работы определяется уникальностью разрабатываемого продукта – речевого корпуса повседневной речи, использование которого может быть востребовано в самых разных областях от решения образовательных и фундаментальных научных задач до внедрения в практику (напр., для отладки телекоммуникационных систем, развития информационных и речевых технологий, решения задач судебной и криминальной фонетики и т. п.).

**Прогнозные предположения о развитии объекта исследования**

На третьем этапе работы (1-2 кварталы 2010 г.) по теме «Звуковое описание грамматических единиц русского языка: информационно-коммуникативный подход» планируется решение следующих задач:

- будет продолжена расшифровка и лингвистическое аннотирование речевого материала. Приоритет в выборе фрагментов для расшифровки будет определяться фонетическим качеством звукозаписей, а также стремлению к получению аннотаций звукозаписей от одного информанта в разных коммуникативных ситуациях.

- продолжатся работы по заполнению базы данных, созданию программ обработки аннотированного материала,

- на материале аннотированных записей будет осуществлено описание особенностей некоторых звуковых реализаций таких грамматических единиц, как морфема и словоформа,

- продожится создание звукового морфемария русского языка повседневного общения: количество анализируемых морфем будет увеличено до 2000.

- будут разработаны дополнительные программные средства, позволяющие автоматизировать процесс обработки аннотаций и исследовать звуковую форму грамматических единиц в схожих и разных условиях (в том числе и в разных коммуникативных сценариях).

- продолжится исследование по выявлению в речи говорящих лексико-грамматических, синтаксических, прагматических особенностей в зависимости от их собственных психологических и социальных характеристик, а также от их социальной роли в коммуникации.