О.Ю. Крючкова

# Задачи информационно-поисковой системы диалектного корпуса и лингвистическое обеспечение поискового механизма

# Общая концепция СарДК

Разрабатывается с сер. 80-х гг. XX в.

1. Диалектный корпус имеет лингво-культурологический характер и создается как полноценный научный источник.
1. Основная задача – представить в корпусе и снабдить грамматической (отчасти и семантической) аннотацией репрезентативное собрание текстов конкретных говоров.
1. Основу корпуса составляют аудио- и видеозаписи, соотнесенные с ними аннотированные символьные расшифровки во вспомогательной транскрипции, а также текстовые и графические материалы историко-культурного характера.

С каждым текстом соотнесены модули справочного характера:
- метаописание текста,
- модуль с биографией диалектоносителя,
- модуль с иллюстративными материалами.


3

С каждым текстом соотнесены модули справочного характера:
- метаописание текста, -
модуль с биографией диалектоносителя,
- модуль с иллюстративными материалами.

# Биография диалектоносителя

Межевая Мария Матвеевна, 1918 г. рожд. По документам год рождения - 1917; так записали, спутав со старшей сестрой, умершей в детстве.

Девичья фамилия – Поликарпова.

Окончила 3 класса, после чего около пятнадцати лет работала в школе уборщицей; затем семь лет – поваром; потом – на сенокосе, возила на быках солому; в войну – на рытье окопов; после войны – скотницей.

Имеет 3 медали.

Воспитывала оставшуюся после смерти сестры племянницу, которая живёт в этом селе.

Было четверо братьев. Два брата погибли на фронте, похоронены в братской могиле; один пришел с фронта без ноги. Никого из братьев и сестер в живых нет.

Была два раза замужем. Первый муж был из старообрядческой общины. Вышла замуж за старообрядца, потому что в селе мало было «мирских» (православных) – всего двадцать семей. Первый муж погиб на фронте. От второго мужа родила восьмимесячного ребёнка. Преждевременные роды произошли из-за того, что пришлось поднимать тяжёлые мешки.

Муж погиб, попав под лошадь.

Сын живёт в этом же селе.

Мария Матвеевна живёт с внучкой, которая работает поваром в столовой.

# Лингвистическая обработка диалектного материала

- расшифровка аудио / видеозаписи диалектной речи;
- составление фонетического комментария к тексту;
- составление словаря диалектных слов и выражений к тексту;
- структурная сегментация текста;
- разметка диалектного текста: жанрово-тематическая, морфологическая, метаразметка.

# Структурная сегментация текстов в СарДК

## Минимальный контекст - абзац

бывало у нас их сколько было/ моленнов-то…/ Коробово/ Доронино/ Кладовкино/ эта/ большая моленна/ Тюревска моленна/ шесть моленнов было в селе-ти// и оне были вот рядом/ вот наша-то была вон тут за горой/ она сейчас на столовой стоит/ это вот тоже их… сломали/ поразвозили/ ходили все рядышком/ все/ и были учёны/ а мы сейчас чё? самоучки// кто нас учит? нет+никто// в церкви-ти вот там учут/ ведь оне учут чего/ оне не правы// у них крещенье какое? еретическо// а у нас християнско//

мы купали в купели или в реке/ или в озере/ или где-то/ в пруду/ вон в речке/ как Исуса Христа/ Иоанн Креститель крестил/ а он [поп в церкви]?/ миропомазание/ то… то брызжет/ то ложки каке-то суёт// разве это закон? разве это божье писание?/ там этого совсем и нет// называется крещение/ а у нас по-ихнему-то крещению/ называется/ если только вы приняли крещение/ святое/ оно святое называется/ то вы после этого крещения называетесь християне/ а оне после свово этого причастия/ у них причастие оно называется/ а не крещения/ называются еретиками// а с еретиком/ тоже есть+написано/ с еретиком/ вот не пить/ и не есть/ и не вкупе богу молиться/ это в моленной вкупе// вот// и ласково слово глаголить нельзя//

# Неоднословные единицы в СарДК

В СарДК все неоднословные лексические единицы маркируются знаком «+» и получают грамматическую характеристику как целые единицы:

| Неоднословная единица | Грамматическая характеристика | Пример |
|---|---|---|
| в аккурат | наречие | матери в+аккурат тридцать восемь годов/ |
| и всё | частица | чёрны рубахи носят и+всё// |
| да и всё | частица | полностью корову накоси/да+и+всё// |
| вот тебе | частица | замуж вышла/ вот+тебе// |
| Красная площадь | идиом | присягу принимал на Красной+площади |
| трудовая книжка | идиом | в трудовую+книжку записали |
| совесть побила | идиом | а потом совесть+побила |

# Неоднословные единицы в СарДК

| Неоднословная единица | Грамматическая характеристика | Пример |
|---|---|---|
| красно солнышко | идиом | красно+… ты …+солнышко |
| Белый Ключ | топоним | там Белый+Ключ |

# Разметка имен собственных в СарДК

- После частеречной пометы вводятся обозначения:
- **name** (имя)
- **sname** (фамилия)
- **part** (отчество)
- **topon** (топоним)
- **mtopon** (микротопоним)
- **nick** (прозвище) **animname** (кличка животного)
- **onim** (остальные имена собственные).

# Жанрово-тематическая разметка в СарДК

- #13@1 бывало у нас их сколько было/ моленнов-то…/ Коробово/ Доронино/ Кладовкино/ эта/ большая моленна/ Тюревска моленна/ шесть моленнов было в селе-ти// и оне были вот рядом/ вот наша-то была вон тут за горой/ она сейчас на столовой стоит/ это вот тоже их… сломали/ поразвозили/ @1#13 #11@1 ходили все рядышком/ все/ и были учёны/ @1#11 #11@2 а мы сейчас чё? самоучки// кто нас учит? нет+никто// в церкви-ти вот там учут/ ведь оне учут чего/ оне не правы// у них крещенье какое? еретическо// а у нас християнско//

- мы купали в купели или в реке/ или в озере/ или где-то/ в пруду/ вон в речке/ как Исуса Христа/ Иоанн Креститель крестил/ а он [поп в церкви]?/ миропомазание/ то… то брызжет/ то ложки каке-то суёт// разве это закон? разве это божье писание?/ там этого совсем и нет// называется крещение/ а у нас по-ихнему-то крещению/ называется/ если только вы приняли крещение/ святое/ оно святое называется/ то вы после этого крещения называетесь християне/ а оне после свово этого причастия/ у них причастие оно называется/ а не крещения/ называются еретиками// @2#11 #11@5 а с еретиком/ тоже есть+написано/ с еретиком/ вот не пить/ и не есть/ и не вкупе богу молиться/ это в моленной вкупе// вот// и ласково слово глаголить нельзя// @5#11

# Перечень тем для разметки в СарДК

1. История жизни.
2. Семья.
3. Дом, быт, домашнее хозяйство, одежда, пища.
4. Трудовая деятельность. Производство. Промыслы.
5. Служба в армии.
6. Пенсия.
7. Односельчане, знакомые, другие.
8. Город, городские.
9. Обряды, обычаи, приметы, праздники. Мифические существа.
10. Здоровье и лечение.
11. Религия.
12. Природа. Погода. Описание местности. Ландшафт.
13. История села.
14. Развлечения, досуг.
15. Государство, власть, политика, общественно-экономическая жизнь.
16. Нормы морали, поведения.
17. Учёба.
18. Происшествия. Случай из жизни (подробный рассказ).
19. Общая оценка жизни.
20. Я-говорящий (самохарактеристика).
21. Язык, речь.

# Перечень жанров для разметки в СарДК

1. Рассказ-повествование.
2. Рассуждение.
3. Описание.
4. Фольклор.
5. Прецедентные тексты (пословицы, поговорки, цитаты, аллюзии, молитвы и др.)
6. Фатическое общение. Метаобщение.

# Морфологическая разметка в СарДК

**Для диалектного корпуса** важно ввести в морфологическую разметку по крайней мере еще две позиции: 1) литературное соответствие каждой размечаемой единице (приводится в круглых скобках после леммы) и 2) помету диалектного своеобразия словоформы (тег – *nstand*).

Морфологическая разметка больших корпусов делается автоматическими программами с ручной корректировкой результатов. Для автоматической разметки в НКРЯ используется анализатор MYSTEM (автор – Илья Валентинович Сегалович). В СарДК применяется одна из версий анализатора ДИАЛИНГ (разработка группы Алексея Викторовича Сокирко).

Пример: **сено-то обрали/ три стога/ да уехали/**

сено{сено(сено)=S,сред,неод=ед,вин}
-то{то(то)=PART}
обрали{обрать(сложить)=V=сов,изъяв,прош,мн=nstand}/
три{три(три)=NUM=вин,неод}
стога{стог(стог)=S,муж,неод=ед,род}/
да{да(да)=CONJ} уехали{уехать(уехать)=V=сов,изъяв,прош,мн}/

При формулировке соответствий обычно приходится обращаться к областным словарям.

# Фрагмент пословно размеченного диалектного текста

ну{ну(ну)=PART} <…> когда{когда(когда)=ADV/CONJ}
стало{стать(стать)=V,сов=изъяв,прош,ед,сред} мне{я(я)=S-
PRO,ед,од=дат} годов{год(год)=S,муж,неод=мн,род=*}
шестнадцать{шестнадцать(шестнадцать)=NUM=им}
или{или(или)=CONJ}
пятнадцать{пятнадцать(пятнадцать)=NUM=им}/
это{это(это)=PART} уже{уже(уже)=ADV}
пошла{пойти(пойти)=V,сов=изъяв,прош,ед,жен}
мода{мода(мода)=S,жен,неод=ед,им}
панталоны{панталоны(панталоны)=S,жен,неод,мн=вин}
белые{белый(белый)=A=мн,вин,неод} с{с(с)=PR}
кружевами{кружева=S,сред,неод,мн=твор}// <…>
и{и(и)=CONJ} нижние{нижний(нижний)=A=мн,вин,неод}
юбки{юбка(юбка)=S,жен,неод=мн,вин}
белые{белый(белый)=A=мн,вин,неод}// <…>
чтобы{чтобы(чтобы)=CONJ}
немного{немного(немного)=ADV}/ из-под{из-под(из-
под)=PR} платьев{платье(платье)=S,сред,неод=мн,род} <…>
видать{видать(видно)=PRAEDIC=*}
было{быть(быть)=V,несов=изъяв,прош,ед,сред}//

# Краткое метаописание в СарДК

- <ФИО полностью=пол=год рождения=образование (неграмотный/ …классов/ техникум и под.)=год записи=область=село=район>
- Например: <Акимов Михаил Родионович=муж=1912=4 класса=1999=Саратовская=Белогорное=Вольский>.

# Полное метаописание в СарДК

1. Сведения об информантах
2. Сведения о времени и месте записи
3. Конкретная ситуация общения
4. Адресаты речи
5. Упоминаемые лица
6. Время описываемых событий
7. Перечень тем (цифровое обозначение)
8. Перечень жанров (цифровое обозначение)

# Папка текста

- 1) звуковой файл (**Miroshin1**);
- 2) файл с расшифровкой во вспомогательной транскрипции (**Miroshin1**);
- 3) файл с морфологически размеченным текстом (**Miroshin1-Razm**);
- 4) файл с метаинформацией (**Miroshin1-Meta**);
- 5) файл с биографией диктора (**Miroshin1-Biogr**);
- 6) файл с фотографиями диктора (**Miroshin)**;
- 7) файл с перечнем раздельнооформленных структурных единиц данного текста (**Miroshin1-Neodnosl_Str**);
- 8) файл с перечнем фразеологизмов, частотных коллокаций, тавтологических и аналитических выражений в тех формах, в которых они представлены в данном тексте (**Miroshin1-Idiom_Slovar**);
- 9) файл с перечнем фразеологизмов, частотных коллокаций, тавтологических и аналитических выражений в начальных формах их элементов (**Miroshin1_Idiom**);
- 10) файл с фонетическим комментарием к тексту (**Miroshin1_Fonetika**);
- 11) файл, представляющий словарное описание диалектных слов и устойчивых сочетаний (**Miroshin1_Slovar**);

# Задачи корпус-менеджера

- выбор подкорпуса конкретного говора;
- получение и сохранение перечня имеющихся в базе данного подкорпуса аудио-, видеоматериалов, текстовых расшифровок;
- воспроизведение речевого фрагмента в любой из форм его хранения;
- получение и сохранение списка информантов в данном подкорпусе и списков всех форм хранения речевого материала, записанного от каждого информанта;

# Задачи корпус-менеджера

- получение и сохранение метаданных о каждом тексте/записи;
- получение и сохранение биографических данных и других видов нелингвистического материала, хранящегося в базе корпуса
- получение и сохранение фонетических комментариев к текстам;

# Задачи корпус-менеджера

- получение списков тем и жанров речевой коммуникации, выделенных в диалектном корпусе;
- создание и сохранение пользовательских корпусов, формируемых по разным параметрам;
- параллельное воспроизведение звуковых записей и их текстовых расшифровок;

## Задачи корпус-менеджера

- получение и сохранение пользователем конкордансов;
- получение и сохранение словаря диалектных слов и выражений и их конкордансов;
- получение и сохранение списков разных видов неоднословных единиц;
- получение статистических данных по разным видам поисковых запросов.

*Спасибо за внимание!*

# Вологодская область. Мегра

Аудиозаписи из материалов
Саратовского диалектного корпуса

## Фольклор
## Этнографические рассказы

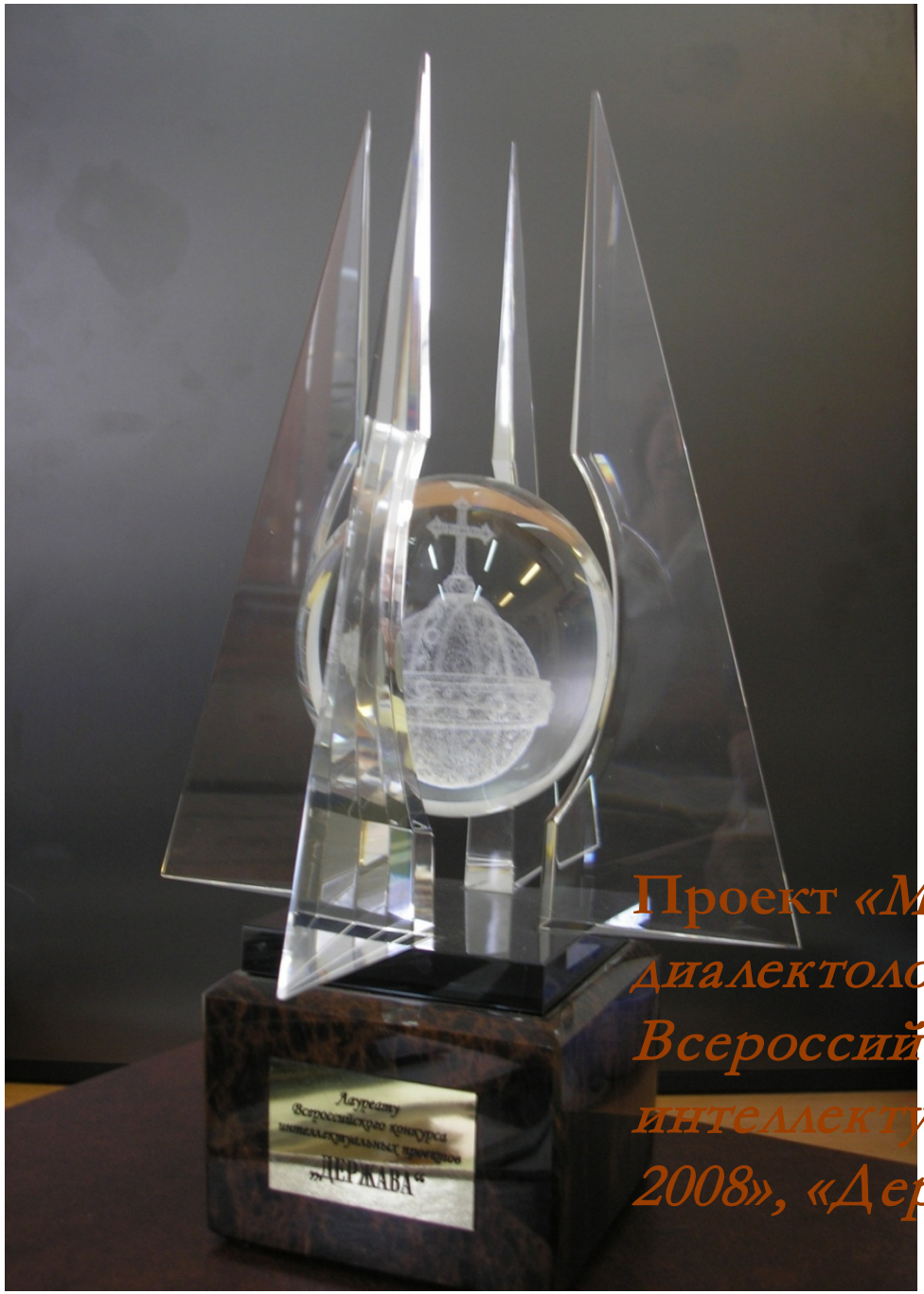
Диалектологи Саратовского государственного университета им. Н.Г.Чернышевского на протяжении нескольких десятилетий ведут активную работу по изучению русских народных говоров.

Аудиозаписи, представленные на этом диске, лишь часть большого материала, собранного саратовскими диалектологами во время экспедиций в Мегру в период с **1975** по **2010** год.

Жители Мегры, от которых сделаны записи:

1. Ерукова Татьяна Степановна
2. Крохина Елена Андреевна
3. Летягина Мария Яковлевна
4. Минкина Павлина Дмитриевна
5. Митина Ольга Ивановна
6. Моськина Мария Ивановна
7. Некипелова Александра Кузьминична
8. Прусская Мария Павловна
9. Тринина Анна Михайловна
10. Фадин Иван Васильевич

Работа выполнена при финансовой поддержке фонда «Русский мир»




ОБЩЕСТВЕННАЯ ПАЛАТА
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФОНД РУССКИЙ МИР

Проект *«Мультимедийный диалектологический корпус»* - лауреат *Всероссийских конкурсов интеллектуальных проектов «Держава-2008», «Держава-2009»*