041-Kruchkova_OJu:_Layout 1 13.05.2008 21:44 Page 268

ТЕКСТОВЫЙ ДИАЛЕКТОЛОГИЧЕСКИЙ КОРПУС КАК МОДЕЛЬ ТРАДИЦИОННОЙ СЕЛЬСКОЙ КОММУНИКАЦИИ[1]

TEXTUAL DIALECT CORPUS AS A MODEL OF TRADITIONAL RURAL COMMUNICATION

Крючкова О.Ю. *(vpks@rambler.ru),* ***Гольдин В.Е.****(goldinve@yandex.ru)*
Саратовский государственный университет им. Н.Г. Чернышевского

В докладе обсуждаются принципы организации и методика построения мультимедийного диалектологического текстового корпуса, представляющего диалект как целостное культурно-коммуникативное образование, моделирующего коммуникацию конкретных речевых коллективов в конкретных социокультурных условиях.

1. Введение

В лингвистике уже разработаны принципы построения текстовых корпусов как коммуникативных моделей (Британский национальный корпус, Национальный корпус русского языка и др.). Наряду с такими корпусами, представляющими функционирование национальных языков в различных сферах общения, необходимы также и корпуса, моделирующие коммуникацию в отдельных языковых сообществах, выделяемых в рамках национального языка. Важнейшими языковыми образованиями такого типа являются диалекты.

Создание диалектных текстовых корпусов имеет особую значимость для диалектологии, ограниченной до последнего времени кругом источников. Основным источником диалектологии до недавнего времени оставался материал, собранный по специальным вопросникам и представленный в картотеках, словарях, атласах. Специфика материала определяла и особый, периферийный, статус диалектологии в составе русистики, значительно сужала круг решаемых диалектологией задач, ограничивая диалектологическую проблематику в основном сферой системно-структурного своеобразия русских народных говоров. Первичный же диалектный материал, имеющийся в различных диалектологических центрах, во-первых, все еще мало доступен широкому кругу исследователей, а во-вторых, в том виде, в котором он обычно существует (без специального аннотирования) не может быть использован с максимальной пользой.

Корпуса диалектной речи, отражающие коммуникацию на диалекте в том или ином конкретном населенном пункте и сохраняющие в машиннообрабатываемой форме значительные массивы связной речи, являются основным источником изучения коммуникативной специфики диалектов. Использование материалов таких корпусов дает возможность не ограничиваться отдельными примерами, а переходить к выявлению общих принципов, тенденций, действующих в диалектной коммуникации (см., напр. работы, выполненные на материале СДК [Гольдин, Крючкова 2006; Балаян 2006; Крючкова, Гольдин, Сдобнова 2007; Крючкова 2007а; Крючкова, Гольдин 2007; Крючкова, Сдобнова, Гольдин 2007; Свешникова 2007а; Свешникова 2007б]), позволяет составить целостное представление о специфике традиционного сельского общения, построить симметричное литературному языку описание диалектной речи.

2. Характеристика имеющихся диалектных корпусов

Мысль о необходимости создания машинного фонда диалектных текстов была высказана еще в 1980-х г. А.С. Гердом и аргументирована В.Е. Гольдиным [Машинный фонд, 1986: 72; Гольдин, 1990]. Однако до сих пор в отечественной и зарубежной лингвистике отсутствуют крупные корпуса диалектной речи, моделирующие коммуникацию на диалекте.

В настоящее время существует ряд корпусов, репрезентирующих отдельные элементы диалектной речи: зарубежные корпуса диалектных текстов (например, Helsinki corpus of English dialects, Kirk's Northern Ireland Transcribed Corpus of Speech (NITCS), IViE (Intonational Variation in English) corpus, BBC Voices); диалектный подкорпус в составе Национального корпуса русского языка (НКРЯ); лексико-грамматическая база данных (ЛГБД) по говору с. Пустоша Шатурского р-на Московской обл., включающая тексты – образцы речи носителей говора [Тер-Аванесова, Крылов, 2006].

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (РФФИ), проект № 06-06-80428-а

041-Kruchkova_OJu:_Layout 1 13.05.2008 21:44 Page 269

Основная задача диалектного подкорпуса НКРЯ заключается в представлении диалекта как специфической территориальной разновидности общенародного языка. Диалектный подкорпус НКРЯ включает текстовые фрагменты различных говоров (напр., говоры Архангельской области, Волгоградской, Рязанской, Вологодской областей), выделенные на тематической основе (быт, гадание, обычаи, жизнь и др.). Методической основой корпуса является последовательное сравнение диалекта с литературным языком (прежде всего в области морфологии и лексики): «в диалектном подкорпусе специально отмечаются отличия от литературного языка» [Летучий, 2005: 215]. Для этой цели при разметке корпуса используется ряд дифференциальных помет, фиксирующих параметры, по которым диалектная текстоформа отличается от соответствующей ей литературной формы (пометы dialmorph, diallex с их последующей конкретизацией); подробнее о диалектном подкорпусе НКРЯ в сопоставлении с Саратовским диалектологическим корпусом см. [Крючкова 2007б].

ЛГБД по говору с. Пустоша отличается ориентацией на один конкретный говор и имеет целью «исчерпывающее описание говора в рамках определенного корпуса текстов» [Тер-Аванесова, Крылов, 2006]. Аннотированный диалектный корпус говора с. Пустоша дает сведения о фонетической, морфологической и лексической специфике текстовых словоформ говора. В корпусе, как и в НКРЯ, использована сравнительная методика описания диалекта, при которой диалектная специфика рассматривается в качестве «отклонений» от литературного аналога.

Таким образом, имеющиеся диалектологические корпуса объединяет дифференциальный подход к диалекту, направленный на выделение элементов, отличающих диалект от литературного языка. Диалектологические корпуса, построенные на дифференциальной основе, демонстрируют территориальное варьирование национального языка, но дают ограниченное представление о традиционном сельском (диалектном) общении как о целостном культурно-коммуникативном феномене.

3. Саратовский диалектологический корпус (СДК)

3.1. Принципы организации корпуса

В Саратовском государственном университете им. Н.Г. Чернышевского создается мультимедийный диалектологический текстовый корпус, целью которого является презентация диалекта как целостного культурно-коммуникативного образования, построение модели традиционного сельского общения на диалекте. СДК базируется на представлении о диалекте как о самодостаточной коммуникативной системе, полно обеспечивающей коммуникативные потребности в условиях традиционного сельского общения. При создании корпуса реализуется недифференциальный подход к диалекту, при котором рассмотрение диалектной речи в ее отношении к литературному языку не является основной задачей.

Каждый отдельный говор в СДК образует самостоятельный подкорпус. В настоящее время корпус включает 3 самостоятельных подкорпуса: подкорпус говора с. Белогорное Вольского района Саратовской области, подкорпус говора с. Земляные Хутора Аткарского района Саратовской области и подкорпус куста сел Мегра Вытегорского района Вологодской области (состав подкорпусов может увеличиваться).

Задача создания модели традиционной сельской коммуникации требует разработки специальной программы, обеспечивающей репрезентативность включаемого в корпус материала. Общим принципом формирования текстовой базы корпуса является принцип полного и адекватного отражения в корпусе специфики диалектного общения. Реализация данного принципа предполагает наполнение каждого подкорпуса разнообразным значительным по объему текстовым материалом, репрезентирующим:

– важнейшие типы диалектной речи (речь бытовую, фольклорную, речь в условиях официального, обрядового общения);

– различные формы речи (диалог, полилог, монолог);

– разнообразную тематику сельского общения;

– социальную дифференциацию носителей говора (по полу, возрасту, профессии, уровню образования).

Для построения коммуникативной модели речевого общения в конкретных условиях жизни конкретного речевого коллектива необходим учет социокультурных условий бытования говора. Решение этой задачи осуществляется путем включения в состав каждого подкорпуса многообразной нелингвистической информации: фотографий, видеоиллюстраций, схем, карт, сведений исторического, социокультурного характера, демографических, этнографических, географических данных. Часть данной информации соотнесена в корпусе с текстовыми модулями, другая часть образует отдельный информационный блок (см. Рис. 1).

041-Kruchkova_OJu:_Layout 1 13.05.2008 21:44 Page 270

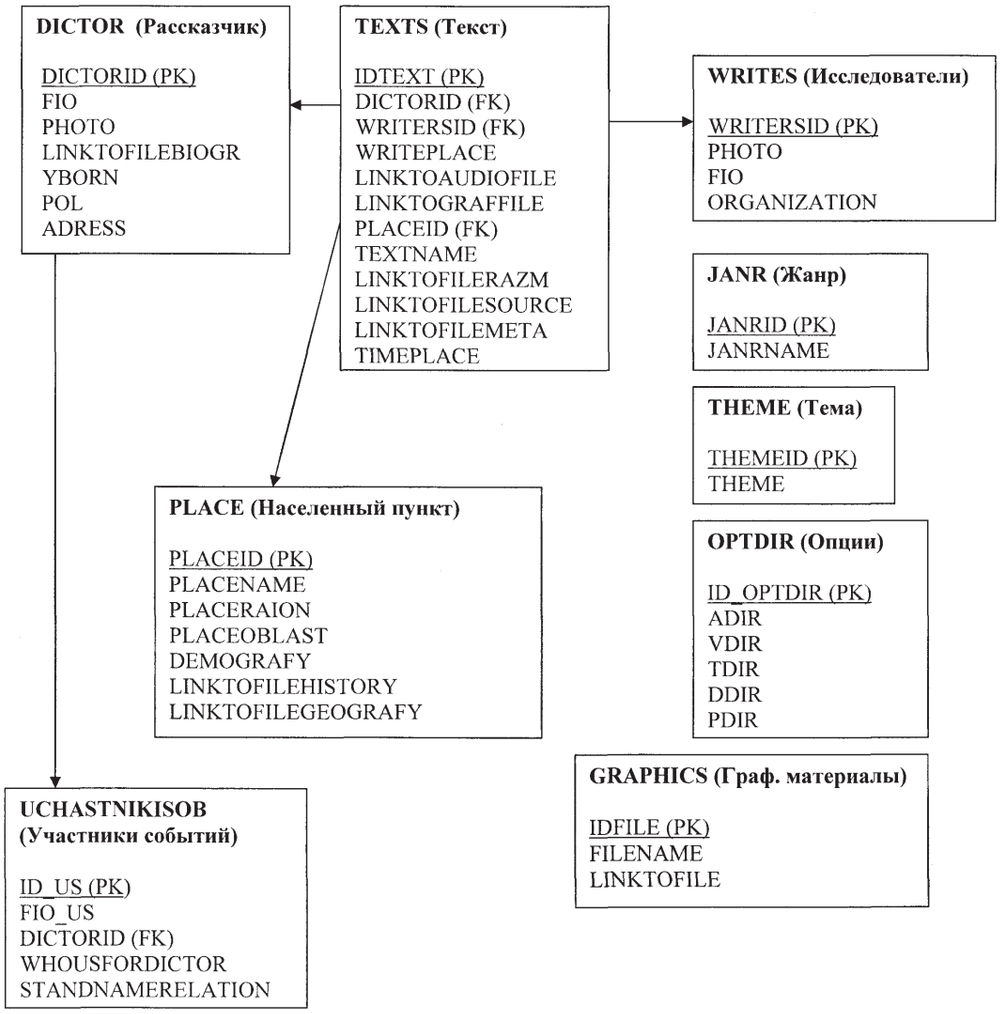

Рис. 1. Информационные блоки СДК

Мультимедийная организация электронной базы корпуса позволяет получать на выходе как отдельные кванты содержащейся в корпусе информации, так и информацию комплексного характера, переходить от одного информационного блока к другому.

3.2. Представление в корпусе диалектной речи

В отношении текстовой (основной) части диалектологического корпуса необходимо решение следующих принципиальных вопросов:

1) способ хранения диалектной речи в базе корпуса;

2) способ членения потока речи;

3) способ символьной записи (расшифровки) устной диалектной речи.

Решения, принятые в СДК:

1) Основная единица базы СДК – «запись»/ «текст» – хранится в трех видах: в виде звукового модуля, в виде текстового модуля с символьной расшифровкой аудиозаписи и в виде текстового модуля с размеченным текстом. От текстовых модулей возможен переход к звуковым модулям и наоборот; программное обеспечение корпуса дает также возможность одновременного воспроизведения аудиозаписи, символьной расшифровки диалектного текста и размеченного текста. В последнее время архив СДК пополняется видеозаписями диалектной коммуникации, значимость которых велика ввиду наиболее полной фиксации коммуникативной ситуации и соотношения между вербальным и невербальным компонентами общения. При наличии видеозаписи звуковой модуль может быть заменен видеомодулем, также программно связанным с текстовыми модулями.

041-Kruchkova_OJu:_Layout 1 13.05.2008 21:44 Page 271

2) Проблема членения речевого потока решена в корпусе в соответствии с принципом максимального приближения модели прототипическому объекту – естественной коммуникации на диалекте. Границы «записи»/«текста» определяются с помощью формального критерия непрерывности общения, так что речевые фрагменты аудио-/видеофайла и текстовых файлов полностью совпадают и соответствуют зафиксированному звукозаписывающей и видеоаппаратурой непрерывному фрагменту общения. В результате применения данного критерия границы звуковых и текстовых модулей не зависят от таких параметров, как смена темы, жанра, формы речи, изменение коммуникативной ситуации и числа участников коммуникации.

3) Расшифровки звучащей диалектной речи приводятся в близкой к орфографической символьной записи, с отражением лексических и грамматических особенностей диалектной речи. Регулярные фонетические явления (например, характер безударного вокализма, диалектные различия в области консонантизма, такие, как диалектное произношение /г/ или /л/) в символьных расшифровках не отображаются. Отражение получают лишь лексикализованные фонетические особенности (типа *топерь, кстить, Рожество, Паска*).

Пример символьной расшифровки диалектного текста в СДК:

вот это у нас образа/ оне называются старообрядческие/ поморские/ вот наша вера какая/ мы не монашки/ мы не какие поповцы/ и не это… вот оне эти вот/ монашки-ти/ оне… как сказать… оне… тоже беспоповцы/ но… у них брак/ и считается за грех/ оне называются безбрачные// оне вот доживают/ до шестьдесят лет/ женщина/ если с мужчиной живёт/ это было раньше/ и у них закон такой/ доживают до шестьдесят лет/ она сносит кануны/ и говорит/ вот/ Ваня там или Вася/ всё/ я больше прекращаю с тобой жизню жить/

Отсутствие фонетической транскрипции в символьных записях восполняется включением в СДК звуковых модулей. Параллелизм текстовых модулей и аудио-/видеомодулей обеспечивает максимальную достоверность фонетической информации, возможность ее наиболее объективного использования в диалектологических исследованиях.

Символьная расшифровка звучащей речи требует также решения целого ряда частных проблем, таких, как возможность и характер использования знаков препинания, обозначение неразобранных фрагментов речи и недоговоренных слов, дифференциация речевых отрезков, принадлежащих диалектологу и диалектоносителю, а также разным диалектоносителям, способ подачи необходимых для понимания текста комментариев. Принятые в связи с данными вопросами решения зафиксированы в специальной инструкции по расшифровке и разметке диалектных текстов СДК (инструкция размещена на сайте www.sarteorlingv.narod.ru).

3.3. Многоуровневая разметка текстов и типы программно обрабатываемой текстовой и метатекстовой информации

Аннотирование текстовой базы корпуса проводится путем многоуровневой лингвистической параметризации диалектных текстов и их метаописания, представляемого в отдельных файлах.

Виды осуществляемой в СДК разметки символьных расшифровок диалектных текстов:

1) пословная лексико-морфологическая разметка;

2) жанровая разметка;

3) тематическая разметка.

Характер разметки определяется реализованным в СДК недифференциальным подходом к диалекту и регулируется следующими положениями:

– все бытующие в говоре языковые формы (совпадающие и не совпадающие с литературными) являются элементами диалектной языковой системы данного говора;

– устное диалектное повествование носит принципиально политематический характер и отличается нечеткими тематико-жанровыми границами.

Названные положения обусловливают особенности проводимой в СДК лексико-морфологической, жанровой и тематической разметки.

Пословная лексико-морфологическая разметка в СДК, полно описывающая все морфологические (классификационные и словоизменительные) признаки текстовой формы и семантику лексических диалектизмов, во многом опирается на принципы, выработанные при разметке текстов в НКРЯ. Разметка проводится с помощью автоматического анализатора с последующим ручным редактированием. Вместе с тем недифференциальный характер СДК обусловил ряд отличий при лексико-морфологической разметке текстов: отказ от дифференциальных помет, применяемых в НКРЯ; характер подачи начальной формы; введение зоны литературных соответствий.

Отказ от дифференциальных помет при пословной лексико-морфологической разметке в СДК не исключает, однако, возможности поиска текстоформ, отличающихся по какому-либо признаку от литературных слово-

041-Kruchkova_OJu:_Layout 1 13.05.2008 21:44 Page 272

форм. Возможность такого поиска обеспечивается специальной маркировкой не соответствующих литературной норме единиц знаком «*», помещаемым в дополнительной зоне разметки.

Начальная форма восстанавливается в СДК на основе конкретной текстоформы, например, для текстоформы *ходилась* приводится начальная форма *ходиться*, для *посклизнулся – посклизнуться*, для *куды – куды* и т.д.

Отказ от дифференциальных помет и текстоориентированная лемматизация словоформ, безусловно, затруднили бы поиск в текстовой базе СДК. Для облегчения поисковых запросов в лексико-морфологическую разметку вводится зона литературных соответствий, идущая в разметке вслед за начальной формой. Литературное соответствие при грамматических, лексических и словообразовательных диалектизмах выполняет также функцию семантической перекодировки. Информативным является и отсутствие литературного соответствия, причиной которого могут быть неясность значения нелитературного слова либо языковая лакуна. В первом случае в зоне литературного соответствия ставится знак вопроса, во втором – тире (прочерк). Ср.: *бедранку* {бедранка(?)=S,жен,неод=ед,вин} (*и он вот под бедранку-то/ залез*); *дак* {дак(–)=PART}, и {и(–)=PART} в функции маркеров конца высказывания.

Примеры пословной лексико-морфологической разметки в СДК:

оне {оне(*они*)=S,мн,од=им=*};

сварывать {сварывать(*сворачивать*)=V,несов=инф=*};

цементовый {цементовый(*цементный*)=A=ед,муж,им=*}.

Ввиду нечеткой выделимости в диалектной коммуникации целостных в жанровом и тематическом отношениях отрезков речи и многочисленных жанрово-тематических переходов и наложений жанровая и тематическая кодировки диалектного текста проводятся на основе предельно обобщенной (не конкретизированной) рубрикации. Элементами жанровой параметризации являются, например, «рассказ-повествование», «рассуждение», «описание», «сказка», «песня», «пословицы, поговорки», а элементами тематической рубрикации – такие темы, как «семья», «обряды. обычаи, приметы», «здоровье и лечение», «религия», «природа», «происшествия».

С каждым текстом связаны 3 модуля метатекстового характера:

1) модуль с метаразметкой текста, элементами которой являются сведения об информантах, о времени и месте записи текста, о конкретной ситуации общения, об адресатах речи, об упоминаемых в тексте лицах, о времени описываемых в тексте событий (до революции; революция и гражданская война; коллективизация; Великая отечественная война; послевоенный советский период; постсоветский период), перечень представленных в тексте тем и перечень жанров;

2) модуль, содержащий биографию информанта, восстанавливаемую по текстовым данным;

3) иллюстративный модуль (фотографии информанта, фотоиллюстрации к данному тексту).

Таким образом, обработка каждого текстового модуля для включения его в СДК завершается формированием папки со следующим набором разноформатных модулей: аудио- / видеомодуль, 4 текстовых модуля (символьная расшифровка аудиозаписи, размеченный диалектный текст, метаописание текста, биография информанта), 1 иллюстративный модуль.

Все элементы пословной лексико-морфологической разметки, жанровой и тематической кодировки диалектных текстов, а также метатекстовая информация являются параметрами поисковых запросов и образуют программно связанное целое.

4. Заключение

Образуя содержательно и программно связанное мультимедийное средство, СДК дает возможность получать комплексную информацию о говоре и условиях его бытования. Электронный диалектологический корпус, моделирующий традиционную сельскую коммуникацию на диалекте, является принципиально новым источником изучения диалектной речи, соответствующим современным требованиям науки о русских народных говорах. Создание СДК и других текстовых диалектологических корпусов будет способствовать приобщению диалектологии к современной научной лингвистической парадигме и построению такой русистики, в которой изучение основных языковых страт (литературной и диалектной речи) находилось бы в необходимой и корректной корреляции.

Список литературы

1. Балаян Э.В. Межконфессиональные отношения в с. Белогорном Вольского района Саратовской области по рассказам местных жителей // Народы Саратовского Поволжья: этнология, этнография, духовная и материальная культура: Материалы межрегиональной научно-практической конференции. Труды Саратовского областного

041-Kruchkova_OJu:_Layout 1 13.05.2008 21:44 Page 273

музея краеведения. Саратов: СОМК. 2006. № 10.

2. Гольдин В.Е., Крючкова О.Ю. Тематическая разметка и тематический анализ диалектного текстового корпуса // Языковая личность – текст – дискурс: теоретические и прикладные аспекты исследования: материалы междунар. научн. конф.: в 2-х ч. Ч.1. Самара: Изд-во «Самарский университет», 2006.

3. Гольдин В.Е. К проекту текстового диалектологического подфонда Машинного фонда русского языка // Доклады Третьей Всероссийской конференции по созданию Машинного фонда русского языка. М., 1990.

4. Крючкова О.Ю. Оценки речи как проявление культурно-языковой идентичности носителей диалекта // Проблема идентичности в современном мире. Саратов: Изд-во Саратовского университета, 2007а.

5. Крючкова О.Ю. Электронный корпус русской диалектной речи и принципы его разметки // Известия Саратовского университета. Новая серия. Филология. Журналистика. Саратов: Изд-во Саратовского университета, 2007б. Т. 7. Вып. 1.

6. Крючкова О.Ю., Гольдин В.Е. Морфологическое своеобразие среднерусской диалектной речи по данным диалектных текстов // Текст и языковая личность: Материалы V Всероссийской научной конференции с международным участием. Томск: Изд-во ЦНТИ, 2007.

7. Крючкова О.Ю., Гольдин В.Е., Сдобнова А.П. Электронный диалектный корпус как новый источник изучения русских народных говоров // Язык и культура в России: состояние и эволюционные процессы: материалы международной научной конференции. Самара: Изд-во «Самарский университет», 2007.

8. Крючкова О.Ю., Сдобнова А.П., Гольдин В.Е. Лексическое своеобразие среднерусской диалектной речи по данным диалектного текстового корпуса // Античный мир и мы: Межвузовский сборник научных трудов. Саратов: Изд-во Саратовского мед. ун-та. 2007. Т.2. № 11.

9. Летучий А.Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., 2005.

10. Машинный фонд русского языка: идеи и суждения. М., 1986.

11. Свешникова Н.В. Модели диминутивного словообразования в русских говорах (функциональный аспект) // Диалектное словообразование, морфемика и морфонология. СПб.: Наука; Вологда: ВГПУ, 2007а.

12. Свешникова Н.В. Явление лексикализации фонетических особенностей в одном из среднерусских говоров // Язык и культура: Материалы Международной научной конференции, посвященной 70-летию профессора Л.В. Савельевой. Петрозаводск: Изд-во КГПУ, 2007б.

13. Тер-Аванесова А.В., Крылов С.А. Лексико-грамматические базы данных как инструмент диалектологического описания // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог 2006». М.: Изд-во РГГУ, 2006.