Как показал анализ, слабые связи в индивидуальном ассоциативном словаре подростка постепенно заменяются сильными, более стандартными, увеличивается доля ядерных лексем в индивидуальном словаре и, таким образом, хорошо прослеживается тенденция сближения ассоциативно-вербальной сети подростка с ассоциативно-вербальной сетью взрослых. Материалы АСШС в целом свидетельствуют о том, что это главная тенденция развития ассоциативно-вербальной сети школьников.

Примечания

1 Работа выполнена при поддержке РГНФ (грант № 05-04-0142а).

2 *Сдобнова А.П.* Индивидуальный ассоциативный словарь школьника // Русский язык сегодня. М., 2004. Вып.3. С.286–295.

3 О структуре и функциональных возможностях АСШС см.: *Гольдин В.Е., Мартьянов А.О., Сдобнова А.П.* Компьютерная версия ассоциативного словаря школьников Саратова и Саратовской области // Русский язык сегодня. М., 2004. Вып.3.

4 В статье имена полей даются прописными буквами, стимулы выделяются полужирным, а реакции – курсивом.

5 Подробнее о типах динамики психологических значений см.: *Гольдин В.Е.* К типологии возрастной динамики ассоциативных полей // Язык. Сознание. Культура. М., 2005.

6 *Колбинева Т.С.* Отказы от реагирования в Ассоциативном словаре школьников Саратова // Язык. Сознание. Культура. М., 2005.

7 См.: *Береснева Н.И., Дубровская Л.А., Овчинникова И.Г.* Ассоциации детей от шести до десяти лет. Пермь, 1995; *Овчинникова И.Г., Береснева Н.И., Дубровская Л.А., Пенягина Е.Б.* Лексикон младшего школьника (характеристика лексического компонента языковой компетенции). Пермь, 2000.

8 *Гуц Е.Н.* Ассоциативный словарь подростка. Омск, 2004.

9 *Караулов Ю.Н., Сорокин Ю.А., Тарасов Е.Ф., Уфимцева Н.В., Черкасова Г.А.* Русский ассоциативный словарь: Ассоциативный тезаурус современного русского языка. М., 1994–1998. Кн.1–6.

10 См.: *Уфимцева Н.В.* Этнический характер, образ себя и языковое сознание русских // Языковое сознание: формирование и функционирование. 2-е изд. М., 2000.

11 Ассоциаты, совпадающие в ядре языкового сознания школьников и взрослых, выделены полужирным.

12 Ср.: *Уфимцева Н.В.* Указ. соч.

УДК 808.2-087

ЭЛЕКТРОННЫЙ КОРПУС РУССКОЙ ДИАЛЕКТНОЙ РЕЧИ И ПРИНЦИПЫ ЕГО РАЗМЕТКИ[1]

О.Ю. Крючкова

Саратовский государственный университет,
кафедра теории, истории языка и прикладной лингвистики
E-mail: pks@rambler.ru

В статье освещаются стратегии создания уникального электронного текстового корпуса русской диалектной речи. Автор рассматривает также принципы разметки диалектных текстов.

Computer-processed Textual Body of Russian Dialect Speech and the Principles of Its Marking

O.Y. Kruchkova

The article deals with the strategies of creating a unique computer-processed textual body of Russian dialect speech. The author also considers the principles of marking dialect texts.

История активного изучения русских народных говоров насчитывает уже более двух веков. За это время была достаточно полно описана территориальная дифференциация русского языка, выделены целостные в языковом отношении территориальные общности различных уровней (группы говоров, наречия, диалектные зоны). Детальное исследование фонетических, грамматических, лексических диалектных особенностей, образующих систему разнообразных по своему характеру междиалектных соответствий (диалектных различий), позволило разработать теорию диалектного языка как особого языкового образования. Вторая половина XX в. была ознаменована такими масштабными диалектологическими проектами, как «Диалектологический атлас русского языка», «Общеславянский лингвистический атлас», «Атлас русских говоров Среднего и Нижнего Поволжья», «Лексический атлас русского языка» (работа над ним продолжается и в настоящее время). Издание атласов и ряда диалектных словарей, среди которых многотомный «Словарь русских народных гово-

ров» (под ред. Ф.П. Филина), создало хорошую источниковую базу для изучения различных языковых особенностей русских народных говоров.

Развитие когнитивно-дискурсивной научной парадигмы в лингвистике привело к формированию нового направления и в изучении русских народных говоров. На рубеже XX–XXI вв. в диалектологии вырабатывается новый подход к пониманию специфики диалекта, согласно которому своеобразие говора не сводится к его структурным особенностям в области фонетики, грамматики и лексики, а проявляется также в строении диалектных текстов, в соотношении различных жанров в составе диалектной коммуникации, в особых приемах раскрытия темы, в когнитивных особенностях диалектной речи, в особой картине мира, реализуемой в общении на диалекте. Задачи и возможности коммуникативной диалектологии были сформулированы в конце 1990-х – начале 2000-х гг. в работах В.Е. Гольдина, томских диалектологов[2]. Постановка новых задач обусловила потребность в новых источниках, необходимость обращения к целостным речевым произведениям на диалекте. В разных диалектологических центрах России (Москве, Саратове, Томске, Благовещенске и др.) ведется активное накопление текстового диалектного материала, который становится основным информационным ресурсом диалектологии.

Развитие современных информационных технологий делает актуальным и возможным создание машинообрабатываемых корпусов диалектных текстов[3]. Работа над созданием электронных корпусов русской диалектной речи ведется в рамках проектов «Национальный корпус русского языка» (Институт русского языка РАН) и «Текстовая репрезентация диалекта как культурно-коммуникативного образования»[4] (Саратовский государственный университет им. Н.Г. Чернышевского). Названные проекты различаются своими целями, принципами организации баз данных, методами обработки (разметки) текстов, включаемых в состав электронного корпуса.

Основная задача диалектного подкорпуса Национального корпуса русского языка (НКРЯ) заключается в представлении диалекта как специфической функциональной разновидности общенародного языка, в демонстрации многообразного территориального варьирования русского языка. В соответствии с этим корпус включает текстовые фрагменты различных говоров; членение корпуса осуществляется на типологической основе (по типам говоров, а не по отдельным говорам). Методической основой корпуса является дифференциальный подход к диалекту, сравнение диалекта с литературным языком (прежде всего в области морфологии и лексики): «в диалектном подкорпусе специально отмечаются отличия от литературного языка»[5]. Для этой цели при разметке корпуса используется ряд дифференциальных помет: dialmorph,type; dialmorph,part; diallex и др. Ср. разметку диалектного текста, подготовленную саратовскими диалектологами для Национального корпуса русского языка: там {там=ADV} было {быть=V=несов, изъяв,прош,ед,сред}/ мало {мало=ADV} зарабатывали {зарабатывать=V=несов,изъяв,прош, мн} / выработок {выработок=S,муж,неод=ед, им} / а {а=CONJ} мы {мы=S,мн,од=им} ***виноваты*** {виноватый=A=кр,мн,им=*dialmorph,type*}/ я {я=S,ед,од=им} оста… всё {весь=A=ед, сред,вин} своё {свой=A=ед,сред,вин} здоровье {здоровье=S,сред,неод=ед,вин} оставила {оставлять=V=сов,изъяв,прош,ед,жен} / это {это=PART}/ пригнали {пригонять=V=сов, изъяв,прош,мн} овец {овца=S,жен,од=мн,род}/ ***овчаркой*** {овчарка=S,жен,од=ед,твор=*diallex*, 'работница, занимающаяся уходом за овцами'} была {быть=V=несов,изъяв,прош,ед,жен}; вот {вот=PART} это {этот=A=ед,сред,им| этот=A=ед,жен,им} вот {вот=PART} тоже {тоже=ADV} моя {мой=A=ед,жен,им} ***насупроть*** {насупроть=ADV=*diallex*,'напротив'}-то {то=PART} вот {вот=PART} эта {этот=A= ед,жен,им} в {в=PR} загородке {загородка=S, жен,неод=ед,пр}-***ти*** {то=PART=ед,жен,пр= *dialmorph,part*}/ картошечка {картошечка=S, жен,неод=ед,им}-то {то=PART}.

Саратовский диалектный текстовый корпус (СарДК), разрабатываемый в Центре изучения народно-речевой культуры Саратовского университета, базируется на ином подходе. Идеологией данного корпуса является представление в нем диалекта как целостной самодостаточной коммуникативной системы в отличие от традиционного дифференциального подхода к диалекту, рассматриваю-

щего диалектную речь в ее отношении к литературному языку. Недифференциальный характер СарДК, его ориентация на диалектную речь одного говора принципиально отличает данный проект от проекта создания диалектного подкорпуса в составе НКРЯ. СарДК создается как модель традиционной сельской коммуникации на диалекте, отражающая речевое общение в конкретных условиях жизни конкретного речевого коллектива. Этот подход определяет структуру корпуса, в котором каждый отдельный говор образует самостоятельный подкорпус[6] и представлен значительным по объему и разнообразным текстовым материалом, соотносимым с многообразной нелингвистической информацией (фотографии, видеоиллюстрации, схемы, карты, сведения исторического, социокультурного характера, демографические, этнографические, географические данные). Единицей хранения в корпусе текстов является «запись» – расшифровка магнитофонной фиксации непрерывного фрагмента общения, приводимая в символьной записи, близкой к орфографической.

Электронная база корпуса позволяет осуществлять запросы, касающиеся грамматических, лексических, словообразовательных языковых явлений, извлекать многообразную нелингвистическую информацию и соотносить ее с лингвистической. С этой целью проводится пословная[7] лексико-морфологическая разметка, а также многоаспектная метаразметка расшифрованных записей диалектной речи. Регулярные фонетические явления (например, характер безударного вокализма) в символьных расшифровках не отображаются, отражение получают лишь лексикализованные фонетические особенности (типа ***топерь, кстить, Рожество, Паска***). Однако в корпусе учитывается значимость фонетической информации; она может быть получена из первоисточника – включаемых в корпус звуковых файлов, от которых возможен переход к символьным записям и наоборот. Такое представление фонетической информации в корпусе делает ее наиболее объективной и пригодной для использования в диалектологических исследованиях.

Лексико-морфологическая разметка в СарДК во многом опирается на принципы, выработанные при разметке текстов в НКРЯ[8]. В СарДК используются, в частности, принятые в НКРЯ правила обозначения классификационных и словоизменительных признаков словоформы. Таким образом, грамматический поиск (поиск с целью получения грамматической информации) в СарДК, так же как и в НКРЯ, может вестись по словоформе, по маске, по начальной форме, по любому из классификационных и словоизменительных признаков.

Вместе с тем недифференциальный характер СарДК 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 связан с задачей представления диалекта в корпусе как целостной самостоятельной коммуникативной системы. Этому подходу соответствует недифференцированное рассмотрение всех бытующих в говоре языковых форм (совпадающих и несовпадающих с литературными) в качестве элементов единой диалектной языковой системы.

Этой же установкой продиктован и принцип подачи начальной формы. Строго говоря, при последовательном воплощении изложенного подхода к диалекту от лемматизации следовало отказаться, так как начальная форма диалектологам нередко доподлинно не известна и при определении леммы неизбежны условные решения. В электронном корпусе начальная форма необходима для удобства поисковых запросов. В связи с этим в СарДК начальная форма восстанавливается на основе *данной* текстоформы, например, для текстоформы ***ходилась*** приводится начальная форма ***ходиться***, для ***посклизнулся – посклизнуться***, для ***куды – куды*** и т.д.

Необходимо отметить, что принятие такого решения не устраняет полностью проблемы начальной формы. В ряде случаев однозначная, пусть даже и условная, лемматизация оказывается все-таки затруднительной. Так, текстоформе ***мозгов*** (*я это/ упала с во-*

зу/ у меня сотрясение мозгов) с одинаковой вероятностью могут быть приписаны начальные формы ед.ч. и мн. ч. – ***мозг*** и ***мозги***; текстоформе ***хуже*** (*мы с тобой сухоньки// а уж старенькей-то будешь/ еще хуже/ вон как у меня одни жилы*) – леммы ***плохой*** и ***худой***. В таких случаях в зоне лемматизации приводятся варианты начальной формы, ср.: ***мозгов*** {мозг=S,муж,неод=мн,род|мозги=S,муж,неод,мн=род}, ***хуже*** {плохой=A/ADV=срав|худой=A/ADV=срав}.

Трудность для принятой в корпусе системы лемматизации представляют диалектные частицы ***ти, ту, те, та*** и под., согласуемые с полнозначными словами по морфологическому, фонетическому или же фоно-морфологическому принципам, ср.: *эти вот как их* ***деревянные-ти палочки-ти***; *как-то вот* ***бутоны-ти*** *вроде видать*; *а она вот* ***с внучатами-ти***; *этот плавает* ***по воде-ти***; *всё это начали строить тут на ихим-то* ***позьме-ти***; *ведь* ***прежде-ти*** *чистили её*. Позиционная незакрепленность диалектных частиц, возможность употребления (иногда преобладающего) в тех же фоно-морфологических условиях частицы ***то*** (ср.: *а так оне…* ***родники-то*** *кто знает; а вон один раз пошли на... на Конец* ***к подруге-то***; *бабки* ***в селе-то***… *уничтожали... малышей-то*; *а вы топерь чай дальние да? дальние* ***сами-то***?) стали основанием для приведения всех диалектных частиц к начальной форме ***то***.

Отказ от дифференциальных помет при разметке включаемых в СарДК диалектных текстов и текстоориентированная лемматизация словоформ, безусловно, затрудняют поиск в текстовой базе. Для облегчения поисковых запросов в лексико-морфологическую разметку вводится зона литературных соответствий, идущая вслед за начальной формой. Ср.: ***оне*** {оне(*они*)=S,мн,од=им}; ***сварывать*** {сварывать(*сворачивать*)=V, несов=инф}; ***цементовый*** {цементовый(*цементный*)=A=ед,муж,им}.

Возможность поиска по литературным соответствиям не только обеспечивает удобство пользования электронной базой диалектных текстов, но и служит отправной точкой для получения интересующей пользователя языковой информации грамматического, лексического и словообразовательного характера. Опора на литературные соответствия позволяет определить сходства и различия диалектных и литературных грамматических форм, лексем, словообразовательных моделей как в плане выражения, так и в семантико-функциональном аспектах, выявить случаи межсистемной (диалектно-литературной) омонимии типа ***овчаркой*** {овчарка(*скотница*)=S,жен,од=ед,твор} (ср.: *пригнали овец/ овчаркой была*); ***всходишь*** {всходить(*приходить в себя*)=V=несов,изъяв,непрош,ед,2-л} (ср.: *долго потом/ всходишь// двое суток в сознание-то не всходила*). Информативным является и отсутствие литературного соответствия, причиной которого может быть неясность значения нелитературного слова либо языковая лакуна. В первом случае в зоне литературного соответствия ставится знак вопроса, во втором – тире (прочерк). Ср.: под {под(под)=PR} ***бедранку*** {бедранка(?)=S,жен,неод=ед,вин} (ср.: *и он вот под бедранку-то/ залез*); ***курбастенький*** {курбастенький(?)=A=ед, муж,им} (ср.: *ну пчела поменьше/ по... там а... тулвище подлиньше/ а трутень курбастенькой такой толстенькой*); ***дак*** {дак(-)=PART}, ***и*** {и(-)=PART} в функции маркеров конца высказывания (ср.: *огуречки/ полить и//*).

Задача создания в корпусе культурно-коммуникативной модели каждого представленного в нем говора определяет важную роль метаразметки при организации базы данных корпуса. Метаразметка осуществляется либо на уровне модулей (соответствующие характеристики приписываются целому тексту), либо на уровне подмодулей (отдельных текстовых фрагментов). Параметрами метаразметки модуля (целого текста) являются сведения об информантах (фамилия, имя, отчество, год рождения, место рождения, родственные связи, образование, род занятий), о времени, месте записи, о конкретной ситуации общения, об адресатах речи, упоминаемых лицах и их отношении к информанту, о хронотопе текста (месте и времени событий в повествовании).

Тематическая разметка, целью которой является моделирование предметной области диалектной коммуникации, делит текст-модуль на подмодули. Предметная специфика диалектного текста (политематичность наряду с функциональной ограниченностью диалектной коммуникации бытовой сферой

общения), а также стремление к сопоставимости тематической структуры диалектных текстов с текстами, представляющими другие типы речи[9], обусловили применение двухуровневой тематической разметки диалектных текстов – широкой и узкой. На уровне широкой тематической разметки выделяются, например, следующие классификационные рубрики: *частная жизнь, дом и домашнее хозяйство, религия, зрелища и развлечения, политика и общественная жизнь, производство.* Узкая разметка отражает тематическую структуру широкой предметной области и служит базой для лексико-семантических и когнитивных исследований диалектной речи. Так, в рамках широкой темы *частная жизнь* выделяются микротемы *родители, дети, другие родственники, брак и семейная жизнь, односельчане.* Тематический блок *дом и домашнее хозяйство* включает подтемы *огород, домашние животные, строения, техника и орудия труда.* Тема *религия* подразделяется на микротемы *православные и иноверцы, религиозные праздники и обряды, религиозный этикет, места религиозных отправлений*[10].

Результатом жанровой кодировки диалектных текстов является выделение жанровых форм речи типа рассказ-повествование, рассуждение, описание, сказка, песня, частушка.

От разных компонентов метаразметки, а также от отдельных текстоформ в СарДК возможен переход к файлам с нелингвистической информацией. Образуя содержательно и программно связанное единое мультимедийное средство, корпус дает возможность получать комплексную информацию о говоре и условиях его бытования, представляет собой новый источник изучения диалектной речи, соответствующий современным требованиям науки о русских народных говорах.

Примечания

1 Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 06-06-80428-а).

2 См., напр.: *Гольдин В.Е.* Теоретические проблемы коммуникативной диалектологии. Саратов, 1997; *Он же.* Изобразительность диалектной речи // Бюллетень фонетического фонда русского языка. №7. Тексты устной речи. СПб., 2000; *Демешкина Т.А.* Теория диалектного высказывания. Аспекты семантики. Томск, 2000; *Иванцова Е.В.* Феномен диалектной языковой личности. Томск, 2002; *Ростова А.Н.* Метатекст как форма экспликации метаязыкового сознания (на материале русских говоров Сибири). Томск, 2000.

3 *Гольдин В.Е.* К проекту текстового диалектологического подфонда Машинного фонда русского языка // Докл. III Всесоюз. конф. по созданию машинного фонда русского языка. М., 1990.

4 Проект поддержан Российским фондом фундаментальных исследований.

5 *Летучий А.Б.* Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С.215.

6 В настоящее время СарДК включает 3 подкорпуса: подкорпус с. Белогорное Вольского района Саратовской области, подкорпус с. Земляные Хутора Аткарского района Саратовской области и подкорпус группы поселений Мегра Вытегорского района Вологодской области.

7 Трудность для последовательного выполнения пословной разметки представляют устойчивые сочетания, смысл которых не выводим непосредственно из значений их компонентов (ср.: *не знай какой работник был* = 'очень хороший работник'). При сохранении пословной разметки устойчивые сочетания отмечаются специальным индексом, отсылающим к имеющемуся в каждом подкорпусе списку таких сочетаний.

8 См.: *Ляшевская О.Н., Плунгян В.А., Сичинава Д.В.* О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005; *Сичинава Д.В.* Обработка текстов с грамматической разметкой: инструкция разметчика // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.

9 О тематической кодировке в НКРЯ см.: *Савчук С.О.* Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.

10 Подробнее о тематической разметке см.: *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. науч. конф.: В 2 ч. Самара, 2006. Ч.I.