УДК 004.934

*Е. А. Жданова*, кандидат филологических наук, доцент
ИжГТУ имени М. Т. Калашникова

# ПРОЕКТ КОРПУСА РУССКИХ ГОВОРОВ УДМУРТИИ

*В статье представлен проект открытого электронного аннотированного полнотекстового корпуса русских говоров Удмуртии, интегрированного в действующую лингвогеографическую информационную систему «Диалект». В статье показаны основные характеристики корпуса в сопоставлении с существующими ресурсами аналогичной направленности, в том числе лингвистическое наполнение, особенности разметки и техническое оснащение, представлены возможности использования проектируемого корпуса в диалектологических исследованиях.*

**Ключевые слова:** лингвистический корпус, русские говоры Удмуртии, диалектный текст, разметка, ЛГИС «Диалект».

*Корпус русских говоров Удмуртии на фоне существующих электронных диалектных корпусов*

На протяжении последнего десятилетия многие диалектологические исследования в Удмуртии проводятся при помощи лингвогеографической информационной системы «Диалект» (ЛГИС «Диалект»), которая на сегодняшний день доступна в Интернете и позволяет хранить диалектный материал в различных формах (паспортизованные лексические данные, собранные по программе ЛАРНГ, транскрибированные записи речи диалектоносителей, аудио- и видеозаписи разговоров с информантами), просматривать (прослушивать) записи, отмечать в текстах диалектные слова, представлять диалектную лексику на масштабируемой лингвистической карте и в виде статей электронного словаря [1]. Таким образом, разработчики ЛГИС «Диалект» вплотную подошли к созданию на основе существующей системы лингвистического корпуса, содержащего диалектные тексты.

По определению В. В. Захарова, лингвистический корпус – это большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных (совокупность текстов), предназначенный для решения конкретных лингвистических задач. [2, с. 3].

В современной диалектологии активно используются возможности корпусной лингвистики. В зависимости от целей создания и предполагаемого направления исследований корпуса содержат текстовый материал русских народных говоров в различных формах, имеют разные типы разметки и объем. Приведем примеры диалектных корпусов, созданных в последние годы.

Наиболее масштабным по охвату говоров и объему материала является Диалектный корпус Национального корпуса русского языка, который был открыт пользователям в 2005 г. и продолжает пополняться и дорабатываться. Этот корпус включает в себя записи диалектной речи (в орфографии, приближенной к стандартной) из различных регионов России. Основная цель Диалектного корпуса НКРЯ – показать своеобразие русских говоров на фоне литературного языка, на что ориентирована и система разметки, разработанная создателями [3]. В корпусе полностью сохранена морфологическая, синтаксическая и лексическая специфика текстов, осуществлена метаразметка и морфологическая разметка, имеются специальные пометы для особенностей диалектной морфологии, собственно диалектные лексемы снабжаются толкованиями [4].

В Центре изучения народно-речевой культуры Саратовского государственного университета имени Н. Г. Чернышевского в 2000-х годах разработан Мультимедийный диалектный корпус (СарДК), содержащий аудиозаписи, сделанные в населенных пунктах Саратовской области, и их расшифровки, произведена лексико-морфологическая разметка и метаразметка. Корпус позволяет осуществлять запросы, касающиеся грамматических, лексических, словообразовательных языковых явлений [5].

С целью сохранения архива, созданного в результате многолетних диалектологических экспедиций, и оптимизации работы с его данными в Томском государственном университете в 2010 г. был разработан проект диалектного Корпуса говоров Среднего Приобья. В числе задач этого проекта – оцифровка существующих транскрибированных записей речи диалектоносителей и создание их машиночитаемых копий, разработка программного обеспечения для метаразметки и автоматизированной лингвистической разметки [6].

На материале вологодских говоров создан корпус диалектных текстов «Жизненный круг», который позволяет проводить исследования диалектного дискурса, этнографические и лингвистические изыскания [7].

В 2013 г. в Школе лингвистики Высшей школы экономики (Москва) совместно с Институтом славянских языков и литератур (Берн, Швейцария) начата работа над проектом «Говор бассейна Устьи. Корпус севернорусской диалектной речи» (http://www.parasolcorpus.org/Pushkino/). Корпус составляют аудиофайлы и транскрибированные тексты, записанные во время экспедиций 2013–2015 гг. в Архангельскую область. Корпус содержит 500 тыс. токенов, имеет морфологическую и синтаксическую разметку. Материалы данного корпуса позволяют проводить исследования фонетики и грамматики указанных говоров [8].

Также в Интернете представлена Электронная библиотека русских народных говоров (http://dialekt.rx5.ru/), которая позволяет прослушивать аудиозаписи речи диалектоносителей из разных регионов России, в том числе содержит 3 аудиозаписи из с. Новогорское Граховского района Удмуртии.

Корпуса диалектного языка разрабатываются и в зарубежных странах. Например, в университете Фрайбурга (Германия) создан корпус русинских говоров (Corpus of Spoken Rusyn: http://www.russinisch.uni-freiburg.de), позволяющий осуществлять поиск слов и словоформ, прослушивать содержащие их аудиофайлы с транскрипцией соответствующих отрывков. Известен Хельсинский корпус диалектов британского английского языка (Helsinki Corpus of British English Dialects: http://www.helsinki.fi/varieng/CoRD/corpora/Dialects/), где представлена орфографическая транскрипция аудиозаписей диалектной речи, сделанных в отдельных районах Англии в 1970-1980-х гг. 20 в.

Во многих странах – в Польше (http://www.dialektologia.uw.edu.pl/index.php), Болгарии (http://corpusbdr.info/), Грузии (http://mygeorgia.ge/ gdc/), в странах Скандинавии (корпус http://www.tekstlab.uio.no/nota/scandiasyn/) и др. – созданы интернет-ресурсы, посвященные местным говорам и содержащие диалектные тексты.

*Состав и техническое оснащение корпуса русских говоров Удмуртии*

В рамках проекта корпуса русских говоров Удмуртии планируется создание интернет-ресурса, где русские говоры Удмуртии впервые будут продемонстрированы в формате электронного аннотированного полнотекстового диалектного корпуса. Материалом для него послужат тексты, представляющие неподготовленную устную речь диалектоносителей из различных населенных пунктов Удмуртской Республики. Эти тексты существуют в виде скан-копий транскрибированных записей устной речи информантов (около 450 тетрадей разного объема из 160 населенных пунктов 20 районов Удмуртии) и аудиозаписей (более 100 аудиокассет и 20 дисков), которые были сделаны студентами и сотрудниками вузов Удмуртии в ходе диалектологических экспедиций в 1970–2000-х гг. и частично внесены в базу данных ЛГИС «Диалект».

В отличие от разрабатываемых другими научными коллективами диалектных корпусов, корпус русских говоров Удмуртии изначально интегрирован в ЛГИС «Диалект», которая дает широкие возможности использования корпусных данных: уже существует модуль для лексикографического представления отмечаемых в текстах корпуса слов и лингвогеографический модуль для визуализации их распространения на лингвистической карте.

Также нужно отметить, что в отличие от многих других региональных корпусных ресурсов, корпус русских говоров Удмуртии будет общедоступным, открытым в Интернете для всех желающих.

Для создания такого интернет-ресурса необходимо «вручную» разметить имеющиеся в разделе «Тексты» ЛГИС «Диалект» скан-копии рукописных диалектных текстов (340 тетрадей) в соответствии с программой Лексического атласа русских народных говоров (ЛАРНГ): для этого в текстах должны быть отмечены и «привязаны» к соответствующим вопросам программы ЛАРНГ все слова, которые являются ответами на данные вопросы. В основу разметки положена именно программа ЛАРНГ, так как в ее вопроснике, включающем более 5000 вопросов, учтено большинство диалектных лексических, словообразовательных и семантических различий, известных ко второй половине 20-го века. При создании корпуса русских говоров Удмуртии упор делается на фиксацию лексических особенностей, поскольку фонетические и грамматические характеристики русских народных говоров изучены в большей степени и систематизация знаний о лексическом составе русского диалектного языка является более актуальной задачей.

Разметка должна сопровождаться «ручной» (не автоматизированной) лемматизацией – указанием начальной формы отмечаемых в текстах слов для удобства осуществления дальнейшего автоматизированного поиска.

Для пополнения базы данных корпуса предполагается ввод, соответствующая разметка и лемматизация не введенных текстовых материалов, содержащихся в тетрадях с записями транскрибированной речи информантов (более 100 тетрадей) и в Хрестоматии по русской диалектологии, изданной сотрудниками Глазовского государственного педагогического института [9], а также создание цифровых копий, ввод и аналогичная разметка сделанных до настоящего времени аудиозаписей бесед с диалектоносителями.

Поскольку в русских говорах Удмуртии, в силу специфических условий их формирования и развития, возможно существование слов, относящихся к тематике, не предусмотренной программой ЛАРНГ, предполагается техническая доработка ЛГИС «Диалект» с целью обеспечить возможность отмечать в текстах и сопровождать соответствующими пометами диалектные слова, выходящие за рамки программы ЛАРНГ, представляющие лексические, словообразовательные, семантические особенности русских говоров Удмуртии.

Помимо этого, создание лингвистического корпуса предполагает разработку корпусного менеджера, позволяющего извлекать из корпуса сведения по тематическим и метахарактеристикам, осуществлять по запросу пользователя выборку из корпуса лексических данных, использовать эту информацию при построении карт для лексического атласа русских говоров Удмуртии и при создании статей для электронного словаря русских говоров Удмуртии.

*Общая характеристика корпуса русских говоров Удмуртии.*

Полученный в результате запланированного исследования интернет-ресурс будет соответствовать всем требованиям, предъявляемым к современному лингвистическому корпусу:

1) Электронная форма. Корпус русских говоров Удмуртии будет состоять из оцифрованных письменных и аудиозаписей устной речи диалектоносителей и будет представлен в открытом доступе в Интернете.

2) Сбалансированность и репрезентативность. Корпус русских говоров Удмуртии будет включать в себя записи живой неподготовленной устной речи диалектоносителей различных возрастов, разного пола, образования и профессии, сделанные более чем в 160 населенных пунктах большинства районов Удмуртии.

3) Унифицированность. Составляющие корпус русских говоров Удмуртии диалектные тексты представляют собой записи, сделанные специалистами-филологами на территории одного региона в современный период (70-е годы 20 в. – 2016 г. 21 в.).

4) Структурированность. Записи речи диалектоносителей хранятся в виде папок с файлами (скан-копиями страниц 1 тетради, которой и соответствует папка) под названием населенного пункта, района республики, года записи и имен информантов (такое расширенное метаописание необходимо, так как во время экспедиции в один населенный пункт обычно заполнялось несколько тетрадей). В папке с записями, которая открывается нажатием на значок «Редактировать», представлен полный список информантов, чья речь приведена в тетради, в виде гиперссылок, которые позволяют просмотреть данные о каждом из них: год и место рождения, образование, вероисповедание и др. Страницы тетради представлены в виде пронумерованного списка скан-копий в формате *jpg, каждая из которых может быть просмотрена в отдельном окне.

5) Значительный объем, обеспечивающий объективность и достоверность данных. Ввод в ЛГИС «Диалект» 450 тетрадей с текстами, включающих обычно несколько десятков страниц, текстов из Хрестоматии по русской диалектологии, а также аудиозаписей (более 100 аудиокассет и 20 дисков) обеспечит количество словоупотреблений (токенов), соответствующее современным представлениям об объеме лингвистического корпуса.

6) Наличие разметки, обеспечивающей получение конкордансов. В корпусе русских говоров Удмуртии на сегодняшний день уже присутствуют основные элементы метаразметки, необходимой для диалектного корпуса (информация о месте, времени, участниках записи). В результате реализации проекта корпус будет снабжен лексической разметкой в соответствии с программой ЛАРНГ, отмеченные слова будут лемматизированы.

В статье О. Ю. Крючковой и В. Е. Гольдина [10], создателей СарДК, представлен ряд специфических критериев, применяемых для оценки корпусов диалектного языка:

1) принципы отбора диалектного материала и критерии репрезентативности диалектного корпуса;

2) принципы членения речевого континуума в корпусе;

3) параметры выдачи текстовых фрагментов;

4) формы представления диалектных текстов в корпусе;

5) виды и правила аннотирования текстовой базы корпуса;

6) параметры метаразметки диалектных текстов;

7) представление в диалектном корпусе нелингвистической информации;

8) оптимальные для диалектологических исследований возможности пользовательских запросов [11].

Введение этих параметров оценки обусловлено своеобразием диалектного материала и особенностями работы с ним. Раскрывая суть этих принципов, авторы, в частности, справедливо отмечают, что диалектный корпус должен давать возможность получать более обширные контексты и даже просматривать входящие в корпус тексты целиком, если этого требует исследование, предпочтительно включение в корпус аудио- и видеозаписей; в статье подчеркнуто, что диалектный корпус требует более тщательной «ручной» разметки, учитывающей все особенности говоров, а не только отличия от литературного языка, особого подхода требует метаразметка; языковой материал в диалектном корпусе, по мнению авторов, должен сопровождаться обширными экстралингвистическими сведениями исторического, этнографического, культурного характера, должна быть указана информация о ситуации записи, упоминаемых в тексте событиях и лицах и т. п.

Применяя предлагаемые параметры к характеристике корпуса русских говоров Удмуртии, можно отметить, что проектируемый корпус будет соответствовать большинству из них. Существенные отличия от концепции О. Е. Крючковой и В. Е. Гольдина заключаются в отсутствии в корпусе русских говоров Удмуртии на первом этапе работы морфологической разметки, что обусловлено иными задачами, стоящими перед разработчиками, а также в невозможности установить для большинства текстов нелингвистическую информацию частного характера, поскольку записи сделаны много лет назад, в то же время в базу данных ЛГИС «Диалект» внесены все сведения о пункте и дате записи, информанте и собирателе, а более конкретные данные о говорящем и ситуации записи (место беседы, присутствие родственников или соседей, состояние здоровья и т. п.) зачастую можно почерпнуть из самого текста, доступ к которому предоставлен пользователю в полном объеме.

*Использование данных корпуса русских говоров Удмуртии в лингвистических исследованиях*

Создание корпуса русских говоров Удмуртии обеспечит возможность применения к диалектному языку нашего региона корпусного метода (предполагающего оперативное получение статистических данных, быстрый поиск материала, представление контекстов, что чрезвычайно важно для изучения лексики, семантики и словообразования диалектного языка). Разработка общедоступного электронного аннотированного корпуса диалектных текстов облегчит исследователям русских народных говоров дос-

туп к материалу русских говоров Удмуртии и даст возможность проводить более широкие по охвату языковых данных системные исследования диалектной лексики, словообразования, семантики. С помощью данного корпуса появится возможность уточнить территорию распространения и период существования интересующих исследователя диалектных явлений. Проектируемый корпус позволит получить статистические данные об употреблении лексических единиц в русских говорах Удмуртии. Только при использовании лингвистического корпуса возможны современные исследования сочетаемости, семантических изменений слов, которые в отношении русских говоров Удмуртии еще не проводились.

Электронный лингвистический корпус русских говоров Удмуртии, помимо общих возможностей, предполагаемых в рамках корпусных исследований, обеспечит также:

– составление уточненных карт лексического атласа русских говоров Удмуртии, на которых помимо данных, собранных по программе ЛАРНГ, будут отображаться и слова, отмеченные текстах корпуса;

– пополнение базы данных Лексического атласа русских народных говоров: методы прямого опроса и тематических бесед с информантами, применяемые собирателями материала для ЛАРНГ, не всегда дают полную и достоверную информацию о лексическом составе русских народных говоров, поэтому привлечение данных из корпуса русских говоров Удмуртии позволит уточнить имеющиеся сведения о лексике русских говоров нашего региона, а в некоторых случаях поможет закрыть лакуны на территории Удмуртской Республики в составе общих карт ЛАРНГ;

– поиск контекстов для составления словаря русских говоров Удмуртии: лексикографический модуль ЛГИС «Диалект» позволяет осуществлять поиск запрашиваемых данных по всему материалу, введенному в систему, таким образом в результате осуществленной разметки слова, отмеченные в корпусе, будут представлены пользователю в контексте как примеры употребления заглавного слова словарной статьи;

– расширение словника словаря русских говоров Удмуртии: отмеченные в корпусе диалектные обозначения, выходящие за пределы программы ЛАРНГ, также будут представлены в виде словарной статьи пользователям ЛГИС «Диалект» при соответствующем запросе;

– применение в учебно-методической работе: созданный электронный ресурс может быть использован при разработке курсов и чтении лекций историко-филологического цикла в вузах, при составлении учебно-методических пособий, при выполнении студентами курсовых и выпускных квалификационных работ, прохождении учебной практики.

Таким образом, создание корпуса русских говоров Удмуртии способствует лексикографическому и лингвогеографическому описанию диалектной лексики нашей республики.

Разработка корпуса русских говоров Удмуртии может стать основой для продолжения изучения диалектного языка этого региона и в других аспектах: наличие доступных в электронном виде транскрибированных и аудиозаписей речи диалектоносителей позволит осуществить в перспективе их морфологическую и синтаксическую разметку, а также отметить фонетические особенности речи информантов, что даст возможность уточнить существующие разрозненные данные о грамматических и фонетических чертах русских говоров Удмуртии, создать соответствующие карты, отражающие диалектное членение анализируемых говоров.

Подчеркнем, что материалом для корпуса служат записи связной речи информантов, которые повествуют о своей жизни, об истории края, о традициях и обычаях местного населения. С этой точки зрения создание ресурса, обеспечивающего доступность подобных материалов, даст информацию для исторических, этнографических, краеведческих исследований.

**Библиографические ссылки**

1. Лингвогеографическая система «Диалект»: история создания, новые возможности, технологические решения, демонстрация данных / В. А. Баранов, Е. А. Жданова, Д. Б. Кожевников, А. А. Белых // Интеллектуальные системы в производстве. – 2013. – № 1 (21). – С. 171–175.

2. *Захаров В. П.* Корпусная лингвистика : учеб.-метод. пособие. – СПб., 2005. – 48 с.

3. *Летучий А. Б.* Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. – М. : Индрик, 2005. – С. 215–232.

4. Национальный корпус русского языка: состав и структура. – URL: http://www.ruscorpora.ru/corpora-structure.html.

5. *Крючкова О. Ю.* Электронный корпус русской диалектной речи и принципы его разметки [Электронный документ]. – URL: http://sarteorlingv.narod.ru/dialekt/ elektr_ korpus.html#ftn0.

6. *Юрина Е. А.* Томский диалектный корпус: в начале пути // Вестник Томского государственного университета. Филология. – 2011. – № 2 (14). – С. 58–63.

7. *Драчева Ю. Н.* Диалектный дискурс в массовой коммуникации // Актуальные проблемы русской диалектологии. К 100-летию издания диалектологической карты русского языка в Европе: тезисы докладов Международной конференции 30 октября – 1 ноября 2015 г. – М. : Институт русского языка им. В. В. Виноградова РАН, 2015. – С. 59–62.

8. *Виняр А. И., Герасименко Е. А.* Исследование правил дистрибуции вариантов постпозитивной частицы *-то* в севернорусском говоре с использованием корпусных методов // Актуальные проблемы русской диалектологии. К 100-летию издания диалектологической карты русского языка в Европе: тезисы докладов Международной конференции 30 октября – 1 ноября 2015 г. – М. : Институт русского языка им. В. В. Виноградова РАН, 2015. – С. 34–37.

9. Хрестоматия по русской диалектологии // авт.-сост.: В.Н. Мартьянова, А. О. Семакина, С. В. Шепелева. – Глазов : ГГПИ, 2014.

10. *Крючкова О. Ю., Гольдин В. Е.* Корпус русской диалектной речи: концепция и параметры оценки [Электронный ресурс]. – URL: http://www.dialog-21.ru/digests/ dialog2011/materials/ru/pdf/36.pdf

11. Там же. С. 360.

* * *

*E. A. Zhdanova*, PhD in Philology, Kalashnikov ISTU

**Science Project of the Linguistic Corpus of Russian Dialects of the Udmurt Republic**

*This article presents the electronic annotated full texted corpus of the Russian dialects of the Udmurt Republic integrated with linguistic-geographical information system "Dialect". The article shows the main characteristics of the corpus versus similar resources including linguistic material, marking features, its technical equipment, and the possibility of using this corpus in dialectological researches.*

**Keywords**: linguistic corpus, Russian dialects of the Udmurt Republic, dialect text, marking, linguistic-geographical information system ”Dialect“.

Получено: 06.12.16