И.Б. Качинская, Д.В. Сичинава

Пополнение Корпуса диалектных текстов в национальном корпусе русского языка

Аннотация: В сообщении дается краткое описание сайтов, которые будут полезны специалистам-диалектологам и лингвистам, интересующимся русской диалектологией. Особое внимание уделено Корпусу диалектных текстов в составе Национального корпуса русского языка. Рассказано о принципах разметки диалектных текстов, о пополнении подкорпуса в 2016 году.

Ключевые слова: русская диалектология, корпусная лингвистика, лексическая семантика, электронные ресурсы

Abstract: The paper briefly describes websites that could be useful for professional dialectologists and for linguists who are interested in Russian dialectology. The Dialectal texts corpus within the Russian National Corpus is in the focus. The main principles of marking up the dialectal texts are presented as well as the 2016 release of the corpus.

Key words: Russian dialectology, corpus linguistics, lexical semantics, electronic resources

Доступ к диалектным текстам до сих пор затруднен даже для специалистов-диалектологов, хотя к настоящему времени опубликовано значительное количество учебников и хрестоматий, содержащих тексты определенного региона (регионов). Как правило, эти учебники и хрестоматии выпускались малыми тиражами в качестве учебных пособий для студентов соответствующих вузов. Сейчас многие диалектологи заняты составлением диалектных словарей. Эти словари опираются на значительные по объему картотеки, часто содержащие не только иллюстрации диалектного слова, но и достаточно развернутые контексты. Во многих университетах собраны также большие фонотеки, расшифрованные лишь в малой части.

В последние десятилетия диалектологи активно осваивают виртуальное пространство. Несколько электронных продуктов имеют учебные цели и предназначены для студентов и школьников: сайт «Школьный диалектологический атлас “Язык русской деревни”» (ИРЯ РАН) (gramota.ru/book/village); сайт «Фонетика русских диалектов» (МГУ имени М.В. Ломоносова) (dialect.philol.msu.ru/index.php).

Постоянно действует (но редко обновляется) сайт «Информационный центр “Русская Диалектология”» (Институт Русского языка им. В.В. Виноградова РАН), созданный «для обмена информацией между различными коллективами россий-

ских и зарубежных диалектологов, обеспечения планомерной и целенаправленной исследовательской и полевой работы в области диалектологии» (www.ruslang.ru/agens.php?id=rus_dialectology). На этом сайте представлены основные центры России, в которых проводится диалектологическая работа.

Постоянно функционируют сайты, на которых размещены электронные версии диалектных словарей, например «Словарь русских народных говоров» (ИЛИ РАН, iling.spb.ru/vocabula/srng/srng.html); «Архангельский областной словарь» (МГУ имени М.В. Ломоносова, www.philol.msu.ru/~dialectology/dictionary/).

В постоянном доступе находится «Акустическая база данных по русским говорам. Диалектная фонетика» (Институт Славяноведения РАН, dialect-phon.ruslang.ru/).

Интерес к прикладной лингвистике, созданию словарей, в том числе диалектных, сопровождается также большой работой по созданию диалектных корпусов.

Обширный аудиоматериал по русским говорам, собранным в различных областях Европейской части России, выложен на сайте Казанского (Приволжского) федерального университета (dialekt.rx5.ru/index.html), ко многим аудиофайлам имеются расшифровки.

Тексты из разных областей представлены в «Лингвогеографической системе “Диалект”» (Ижевск, Удмуртский университет, io.udsu.ru/dl/common.logon). Эта система особенно интересна возможностью работы с интерактивными географическими картами.

Ежегодно пополняется имеющий аудиосопровождение мультимедийный корпус диалектных текстов «Говор бассейна Устьи. Корпус севернорусской диалектной речи» («Ustja River Basin Corpus Query interface»; ВШЭ, Москва и Берн, Швейцария, www.slavist.de/Pushkino/login.php).

«Электронные базы данных по русским народным говорам» (С.А. Крылов и А.В. Тер-Аванесова) включают тексты, записанные в деревнях Харовского р-на Вологодской обл. и Шатурского р-на Московской обл. (starling.rinet.ru/cgi-bin/main.cgi?root=ruscorpora&encoding=utf-rus).

С материалами по русским говорам Кубани можно познакомиться на сайте «Региональная этнолингвистика» (www.ethnolex.ru/).

Большое внимание народным говорам в корпусной лингвистике уделяется за рубежом (в Польше, Чехии, Словакии, Словении, Германии и др.).

Ссылки на некоторые корпуса, имеющие постоянный электронный адрес и содержащие диалектные материалы, даны на сайте НКРЯ www.ruscorpora.ru/corpora-other.html (Другие корпуса: Диалектные корпуса русского языка).

Далеко не все корпуса и другие электронные ресурсы предназначены для широкого пользователя и имеют выход в Интернет.

По материалам трех русских говоров (двух южных, расположенных в Саратовской обл., и одного северного, расположенного в Вологодской обл.) созданы Диалектные корпуса в Центре изучения народно-речевой культуры Саратовского государственного университета им. Н.Г. Чернышевского (руководители – проф. В.Е. Гольдин и проф. О.Ю. Крючкова).

Есть несколько выпусков «Тамбовской фонохрестоматии» (Тамбовский университет), в которой расшифрованные тексты даны в сопровождении аудиоматериалов, имеется карта области, разделенная на районы, включена система Поиска, т. е. по сути эта фонохрестоматия является корпусом. К сожалению, этот электронный ресурс не имеет постоянного адреса в Интернете и распространяется на дисках.

Материал более чем из ста говоров Архангельской области содержится в «Электронной картотеке “Архангельского областного словаря”» (МГУ имени М.В. Ломоносова). Общий объем ее превысил 2 млн «карточек»; как видно из названия, этот электронный продукт имеет жесткую лексикографическую направленность. Система поиска в «Электронной картотеке» существует, но еще недостаточно отработана.

Полезным ресурсом, содержащим ссылки на электронные версии диалектных словарей, на диалектологические карты, атласы, оцифрованные книги и статьи по русской диалектологии, является сайт dlibrary.livejournal.com/561.html/ (дата обращения: 07.12.2016). К сожалению, многие ссылки здесь даны на временно размещенные объекты и уже устарели.

Во всех этих корпусах и других электронных ресурсах по-разному решаются возникающие перед всеми диалектологами проблемы отражения фонетики, грамматики, лексики; часто они жестко направлены на исследования, традиционно проводимые лингвистическими кафедрами соответствующих вузов.

Корпус диалектных текстов входит в состав Национального корпуса русского языка (НКРЯ) и размещен в открытом доступе **(**www.ruscorpora.ru/search-dialect.html). Работа над ним была поддержана фондом РГНФ (2006–2008 гг. № 06-04-13818в: «Создание корпуса диалектных и фольклорных текстов на русском языке», рук. В.М. Живов; 2009–2010 гг. № 09-04-12159в: «Корпус диалектных текстов Национального корпуса русского языка: Грамматическая, фонетическая и метатекстовая разметка. Новый стандарт подачи», рук. В.М. Живов); 2014–2016 гг. № 14-04-12012в «Корпус диалектных текстов Национального корпуса русского языка. Пополнение и разметка» (рук. Д.В. Сичинава).

Коллектив разработчиков приносит Фонду благодарность за поддержку проектов.

Результатом первого (пилотного) Проекта было само создание Диалектного подкорпуса в составе НКРЯ. Результатом следующих Проектов стала разработка нового стандарта подачи текстов и их обработки, благодаря чему появился новый системный продукт «Рабочее место диалектолога», в котором осуществляется разметка диалектных текстов на всех уровнях: метатекстовом и грамматическом; оказалось возможным представить текст с диакритиками (ударениями) в фонетической транскрипции двух видов: «начальной» и «облегченной», унифицированной; была значительно усовершенствована грамматическая разметка; пополнился банк диалектных текстов. Многие из размеченных текстов уже выставлены на сайте, еще большее их количество готовится к выставке в ближайшее время.

«Корпус диалектных текстов» НКРЯ предполагает включение любых диалектных текстов на русском языке, записанных как на территории исконного проживания русского населения (Европейская часть России), так и на территориях раннего заселения (Русский Север), позднего заселения (Сибирь, Дальний Восток, Дон, Нижнее Поволжье) и миграций (говоры старообрядцев / протестантов Латгалии, Азербайджана, Румынии, Австралии, Канады, Америки и т. д.). Тексты предоставляются диалектологами, ведущими полевую работу. Это могут быть записи из полевых тетрадей или аудиорасшифровок, уже введенные в компьютер; тексты из опубликованных хрестоматий, предоставленные в компьютерном варианте.

Для каждого диалектного текста осуществляется *метаразметка*, которая содержит три уровня:

1) адрес-сопровождение (включает все необходимые «паспортные» данные текста: кем, когда, где, от кого записан текст, публиковался ли, где хранится и проч.);

2) фонетический уровень. Отмечаются некоторые особенности фонетики говора в области вокализма и консонантизма. В банке текстов, переданных для выставления на сайт Национального корпуса, многие тексты оказались поданы вовсе не в транскрипции, а в орфографии, и в «фонетической метаразметке» существует помета «Орфографизированная ли запись?». Помета об «орфографизированной записи» будет постоянно присутствовать и на сайте, чтобы пользователь не пытался делать выводы о фонетических особенностях говора на основе текстов, которые первоначально представлены не в транскрипции;

3) диалектная текстовая метаразметка содержит 3 подуровня: жанр (тип) текста; тематика текста; место и время описываемых событий.

Жанр (тип) текста делится на 4 категории: 1) устные нефольклорные тексты; 2) письменные нефольклорные тексты; 3) устные фольклорные тексты; 4) письменные фольклорные тексты. Пока предпочтение отдается устным нефольклорным текстам, хотя и в них могут содержаться фольклорные элементы, которые отмечаются по мере встречаемости (в устном рассказе могут оказаться и колыбельные песни, и частушки, и пословицы, поговорки, загадки и проч.). В Банке диалектных текстов есть письменные фольклорные тексты (тетради заговоров и «песенники», записанные самими носителями) и значительное количество письменных нефольклорных текстов (письма, мемуары, дневники и проч.).

Место и *время* описываемых событий в основном совпадает с тем, как этот же раздел представлен в Основном корпусе НКРЯ.

Тематическая разметка, как нам кажется, требует значительной переработки для упрощения Поиска.

Текст в Диалектном подкорпусе может выдаваться в виде «нарезки» – при поиске лексемы или конкретной грамматической характеристики пользователь получает строку с искомым словом + по два предложения слева и справа (как это устроено в Основном корпусе). Но может, по запросу, получить целый текст. Изначально предполагалось, что текст Пользователю будет выдаваться в двух вариантах: в первоначальной записи – так, как он был подан в Диалектный корпус (в фонетической транскрипции или в орфографизированном виде, так называемый Текст-1), и в унифицированном виде (в «облегченной транскрипции», сохраняющей ударения, но более удобной для цитирования, так называемый Текст-2). По техническим причинам Текст-2 в ближайшее время на сайте показываться не будет. Но там, где тексты первоначально были поданы в орфографизированной записи, Текст-1 и Текст-2 практически совпали.

Все тексты представлены со «снятой омонимией», т. е. с полной грамматической разметкой. Разметка осуществляется в том числе с указанием диалектных особенностей. В отличие от разметки, поиск диалектных особенностей еще недостаточно отработан, но представляется, что он может осуществляться по обычному запросу: Род. или Дат.-Предл. падеж существительных 1 скл., *ся*-глаголы и т. д. или по точным формам типа *ихний, евонный, егонный, ейный*. Возможно, он может осуществляться с помощью той же грамматической таблицы, учитывающей диалектную грамматику, которая появляется в «Рабочем месте диалектолога» при разметке.

В 2016 г. подкорпус пополнился текстами, записанными в Архангельской, Тверской, Тамбовской, Тюменской, Самарской, Волгоградской областях, в Ставрополье, Забайкалье и некоторых других местах. В ближайшее время на сайте появятся тексты из Вологодской, Томской, Пермской областей, говоры Среднего Поволжья, Башкирии, русских молокан Азербайджана. Некоторые расшифровки (из Тверской обл.) дополнены звукорядом. В самом скором будущем появится возможность и видеосопровождения.

Мы надеемся, что со временем этот корпус станет репрезентативным собранием диалектных текстов и будет достаточно востребован пользователями.

Свободное предоставление в Интернете текстов русских народных говоров, их грамматическая, семантическая и метатекстовая характеристика позволят специалистам-диалектологам, другими лингвистам и нелингвистам, филологам, историкам, культурологам, этнографам – всем, кто интересуется народным русским словом, обращаться к корпусу в самых разных целях: примеры из текстов и сами тексты могут выступать в качестве справочного материала, использоваться как материалы для научной и педагогической работы, для демонстрации этнографических, этнокультурных традиций, особенностей русского менталитета.

Сведения об авторах:
Ирина Борисовна Качинская,
канд. филол. наук
мл. науч. сотрудник
филологический факультет
МГУ имени М.В. Ломоносова

Дмитрий Владимирович Сичинава,
канд. филол. наук
ст. научный сотрудник
Институт русского языка им. В.В. Виноградова РАН,

Irina B. Kachinskaya,
PhD
Research Associate
Philological Faculty
Lomonosov Moscow State University

DmitryV. Sichinava,
PhD
Senior Researcher
Vinogradov Institute of the Russian Language, RAS