УДК 81'282 и 81'33
DOI: 10.17223/22274200/11/5

**И.Б. Качинская, Д.В. Сичинава**

# О КОРПУСЕ ДИАЛЕКТНЫХ ТЕКСТОВ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА[1]

*Статья посвящена сегодняшнему состоянию Корпуса диалектных текстов в составе Национального корпуса русского языка. В 2016 г. подкорпус пополнился новыми материалами. Новый стандарт подачи предполагает установку на сохранение транскрипционной записи, ударений, возможность работать не только с фрагментами текстов, но и с целыми текстами. Описываются принципы грамматической, семантической разметки и метаразметки диалектных текстов, принципы поиска нужных материалов на сайте.*
Ключевые слова: *русская диалектология, корпусная лингвистика, лексическая семантика, электронные ресурсы, морфологическая разметка.*

**1. Начало большого пути.** В последние десятилетия наряду с продолжающейся разработкой диалектной лексикографии и лингвогеографии появилась новая возможность изучения русских говоров, связанная с развитием корпусной лингвистики. Созданы электронные ресурсы, имеющие постоянные адреса в Интернете (см. работы [1–9] и некот. др.). В этих корпусах по-разному решаются возникающие перед всеми диалектологами проблемы отражения фонетики, грамматики, лексики; часто они 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 по адресу http://www.ruscorpora.ru/search-dialect.html. Диалектный подкорпус доступен для поиска с декабря 2006 г. За это время поменялся коллектив разработчиков, изменились и принципы представления материала. В пилотном проекте все диалектные тексты были поданы в орфографизированном виде [10, 11], осуществлялась метатекстовая и грамматическая разметка, однако диалектные грамматические особенности отмечались непоследовательно и не всегда верно [12]. Новый коллектив разработал новый стандарт подачи диалектных

[1] Работа осуществляется благодаря поддержке Российского гуманитарного научного фонда, проект № 14-04-12012в «Корпус диалектных текстов Национального корпуса русского языка. Пополнение и разметка».

текстов. Прежде всего делалась установка на сохранение транскрипционной записи с обязательным ударением (при условии, что изначально тексты были поданы в транскрипции и с ударениями). Была создана среда «Рабочее место диалектолога» (доступная для скачивания на сайте http://mitrius.ru/dialect.html вместе с инструкцией), в которой разметка диалектных текстов осуществляется на всех уровнях: метатекстовом и грамматическом. Новый стандарт подачи позволил предоставить пользователю возможность работать не только с фрагментами текстов (как в других подкорпусах НКРЯ), но и с целыми текстами [13–15].

**2. Метаразметка** содержит три уровня: 1) адрес-сопровождение; 2) фонетический уровень; 3) жанрово-тематический.

*Адрес-сопровождение.* При каждом тексте указываются следующие параметры:

1) название текста;

2) лицо, предоставившее текст (как правило, это исследователь-диалектолог): имя, отчество, фамилия; научное звание, должность, место работы; информация для связи (адрес, телефон, e-mail);

3) кем производилась запись;

4) место записи (область, район, населенный пункт);

5) год записи;

6) сведения об информанте: имя, отчество, фамилия; возраст / год рождения; место рождения; образование; профессия / род занятий; другие сведения (конфессия и пр.);

7) где публиковался текст (если публиковался);

8) место хранения записи (научное / учебное учреждение, факультет, кафедра, фонд, личный архив).

Полный паспорт заполняется для всех текстов.

*Фонетическая метаразметка.* Пока отмечаются лишь немногие фонетические диалектные особенности:

1) в области ударного вокализма: позиционные чередования гласных после мягких согласных на уровне «старого ятя» и /а/;

2) в области безударного вокализма: оканье и аканье (включая указания на полное / неполное оканье или диссимилятивное аканье);

3) в области консонантизма: /г/ взрывной или фрикативный; цоканье.

Для публикации на сайте Национального корпуса многие тексты оказались поданы вовсе не в транскрипции, а в орфографии, и в

«фонетическом» блоке метаразметки существует помета *Орфографизированная ли запись?* Возможно выбрать подкорпус только орфографизированных или только неорфографизированных текстов, а также получить эту информацию о каждом конкретном тексте. Это сделано для того, чтобы пользователь не пытался делать выводы о фонетических особенностях говора на основе текстов, которые изначально представлены не в транскрипции.

*Текстовая метаразметка* содержит три подуровня: жанр (тип) текста; тематику; место и время описываемых событий.

*Жанр (тип) текста* делится на четыре категории: устные нефольклорные тексты; устные фольклорные; письменные нефольклорные; письменные фольклорные. Пока предпочтение отдается устным нефольклорным текстам, хотя и там могут содержаться элементы фольклора (в устный рассказ попадают колыбельные песни и частушки, пословицы и поговорки, загадки и проч.). В Банке диалектных текстов уже есть как письменные фольклорные тексты (например, «песенники» или заговоры, записанные самими носителями), так и значительное количество письменных нефольклорных (письма, мемуары, дневники и др.).

*Тематическая разметка* достаточно развернута, первоначально она создавалась на базе вопросников ЛАРНГ. Опыт работы, однако, показал, что она требует значительной переработки для упрощения поиска. Необходимо, с одной стороны, убирать тематические повторы (*Жизнь // Быт* или *Погребально-поминальные обряды, похороны // Смерть*), с другой – конкретику: например, в разделе *Духовная культура, религия* оставить *Народный календарь и календарная обрядность* и убрать *Рождество. Святки. Крещение. Масленица. Пасха. Троица. Посты*, так как деление это оказалось одновременно и слишком дробным, и совершенно недостаточным.

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

Для того чтобы работать в Корпусе диалектных текстов, необходимо: 1) выйти на сайт http://ruscorpora.ru/search-dialect.html; 2) в правом верхнем углу нажать кнопку «Задать подкорпус» – и далее работать именно в нем (http://ruscorpora.ru/mycorpora-dialect.html). Появляются цветные надписи «Основные параметры текста», «Жанр и тематика», «Фонетика», т.е. подкорпус может быть задан исходя из описанных выше параметров метаинформации.

Возможно также выбрать тексты только в орфографической записи и/или только тексты с аудиозаписью. Параметры метаинформации можно задавать при помощи специального окна выбора, для ряда параметров предусмотрена простановка «галочек» или выбора «кнопки».

При выборе параметров подкорпуса дается список входящих в него текстов. Нажав на название текста, можно увидеть его метаинформацию («паспорт»), а для текстов с аудиозаписью (пока их 15) – прослушать и при необходимости скачать эту запись. Нажав на стрелки в конце названия текста, можно открыть текст полностью.

**3. Грамматическая разметка**

Одним из основных принципов Национального корпуса русского языка является обязательность грамматической разметки.

Еще до работы в основной программе тексты необходимо предварительно подготовить: отделить высказывания информанта от высказываний собирателя (квадратными скобками); разделить словоформы для их корректного распознавания грамматическим парсером: разделить фонетические словоформы на «грамматические» (например *г-дому-то* заменить на *г дому -то*); перевести поданную держателем текстов транскрипцию (иногда изготовленную в самодельных шрифтах, в том числе с использованием сразу нескольких шрифтов) в единый шрифт юникода; перевести текст в формат txt в кодировке UTF-8. При открытии подготовленного текста в «Рабочем месте» автоматически срабатывают детранскрипторы: детранскриптор-1 переводит Текст-1 в Текст-2 (первоначальную транскрипцию в унифицированную), детранскриптор-2 переводит Текст-2 в Текст-3 – орфографический «подстрочник», необходимый для работы стандартного грамматического анализатора. Текст-3 необходимо вручную довести до уровня орфографии, хотя и здесь предусмотрена некоторая помощь размечающему текст диалектологу: специальной программой осуществляется проверка стандартной орфографии, цветной меткой помечаются слова, с орфографией не совпадающие. Тексты 1, 2 и 3 выровнены, как в Параллельном корпусе НКРЯ. После обработки Текста-3 осуществляется стандартная автоматическая грамматическая разметка текста, после чего требуется не только устранить грамматическую омонимию (как в Основном корпусе), но и, ориентируясь на Текст-2, отметить диалектные грамматические особенности тех лексем, где эти особенности встретились. То есть

поверх стандартной грамматической разметки в диалектных текстах предусмотрена возможность отмечать диалектные грамматические особенности лексемы. Для этого в «Рабочее место» внедрены грамматические таблицы по каждой из пяти изменяемых частей речи (глагол, существительное, прилагательное, местоимение, числительное). Если необходимо указать особые грамматические характеристики, то в специальной таблице «Диалектные особенности» нужно выбрать соответствующую грамматическую категорию и пометить ее как «диалектную». Для каждой изменяемой части речи предусмотрена возможность указывать диалектные особенности на следующих уровнях:

**Глагол**: *основа – флексия – суффикс – форма – вид – переходность – возвратность – время*.

Так, например, нестандартные для литературного языка *инфинитивы* отмечаются либо указанием на диалектный суффикс (*пекчи*), либо диалектную основу (*трать, жмать*), диалектную форму (*идтить*). То же с *императивом*, где диалектные особенности могут проявляться либо на уровне суффикса (*посодь*), либо на уровне формы (*доедь, ехай*; как известно, императив от этой основы в литературном языке вообще затруднен). Особенности *спряжения* можно указывать, например, следующим образом:

общее спряжение (*любют*): наст. вр. 3 л. мн. ч. + диал. флексия;

3-е спряжение (*они гулят*): наст. вр. 3 л. мн. ч. + диал. флексия;

Таким образом, в *диалектные особенности* на уровне *флексии* попадут самые разные случаи: ударные окончания без перехода /е/ > /о/ (*идо́ш*), конечное *-ть* в 3 л. (*растёть*), формы без *-т* (*идё*), общее спряж. (*смотрют*), 3-е спряж. (*играт*) и проч. Думается, что специалисту в этом достаточно легко разобраться, особенно при возможности учитывать географические фильтры.

Особые **деепричастия** отмечаются либо на уровне суффикса (*выпимши / выпилши / выпитши*), либо на уровне формы (*ушодчи*). Перфектное значение таких деепричастий может быть отмечено как *особое время*.

**Существительное**: *основа – флексия – суффикс – форма – одушевленность – род – число – падеж – склонение.*

**Прилагательное**: *основа – флексия – суффикс – форма – стяженность – склонение, окончание – полнота / краткость.*

Как это принято в лексикографической практике, в литературном языке начальной формой всех адъективов по умолчанию является И. ед. муж. рода с безударными флексиями *-ий/-ый* (*новый, синий, всякий*) и *-ой* в случаях ударной флексии (*больно́й, большо́й, тако́й*). Это уже сложившаяся традиция, хоть она и противоречит фонемному принципу, определяемому по сильной позиции. В таком случае «диалектной» формой для адъективов должны считаться безударные окончания *-ой/-ей*, встречающиеся в северных говорах с полным оканьем (*ста́рой, си́ней*), или ударные окончания *-ый* и *-ей*, характерные для западной зоны (*золотэ́й, золоты́й, какэ́й, какей́*). При нестандартной сравнительной степени прилагательных / наречий (*лучшее, послабже, первеющий, широкящей* и проч.) помимо указания на *диал. суффикс* и *диал. основу* отмечается *диалектная форма*.

**Местоимение**: *основа – флексия – суффикс – форма – склонение.*

В формах с начальным /j/ (*jон, jона, jоны, jих, jим, jетот* и проч.) отмечается диалектная основа, хотя эти же формы могут рассматриваться и как фонетические варианты и в таком случае вовсе не отмечаться. То же касается соотношения начального гласного / согласного *j*/*н* с предлогом и без предлога: *с им, с jим* = с ним; *доволен ней, ним = ей/ею, им.* Подобные примеры можно искать на соотношение предлог + личное местоимение 3 л.

Достаточно часто трудно / невозможно разделить основу и флексию: местоимения типа *jо* или *jу* (*её = она,* 3 л. ж.р. Вин. ед.), *тэй / тый парень* (= *тот*), *оне, оны* (= *они*). В этом случае отмечается *диалектная форма.* То же при наличии окончаний у притяжательных местоимений 3 л. (*его*, *её, их*): *ихной, ихний, ихой, евонной* и др.

**Числительное**: *основа – суффикс – флексия – форма.*

В неясных случаях синтаксическую межчастеречную омонимию предлагается оставлять, например: *тоже* – част / союз; *ну* – межд / част; *вот* – част / нареч. и т.д.

Все тексты в Диалектном подкорпусе представлены со «снятой омонимией», т.е. с полной грамматической разметкой, в том числе с указанием диалектных особенностей.

Сейчас на сайте http://ruscorpora.ru/search-dialect.html доступен поиск по корпусу объёмом 300 тысяч словоупотреблений. Поиск, как и во всем Национальном корпусе русского языка, возможен в двух режимах: поиск точных форм (в соответствии с Текстом-1, т.е. в орфографизированной или фонетической записи в зависимости от

текста в его первоначальной подаче) и лексико-грамматический поиск. В первом случае задаются конкретные словоформы или их сочетания, во втором – лексемы, наборы грамматических характеристик и семантические толкования и их сочетания (до 10 слов). Грамматические характеристики могут задаваться при помощи специального окна выбора, остальные вводятся вручную. Для каждого из слов, заданных в лексико-грамматическом поиске, может быть задан один или более из перечисленных выше параметров. Таким образом, можно найти любое слово, начинающееся с *под-*, любое существительное 1-го склонения, любое слово, толкуемое как «вот» (при наличии толкования). Кроме того, имеются дополнительные параметры, связанные с пунктуацией, с концом и началом фразы.

В поисковой выдаче доступна та же функциональность, что при выборе подкорпуса (открытие метаинформации и текста целиком, а также имеющейся аудиозаписи). При нажатии на конкретное слово демонстрируются его начальная форма, грамматические характеристики и семантическое толкование (если оно дано). Можно менять параметры выдачи: число контекстов на одной странице, сортировка в хронологическом порядке и т. д.

**4. Семантика**

Информация по семантике предоставлена на двух уровнях: метаразметки и пословной разметки текста. Тематика текста определяется на уровне метаразметки, при этом рекомендуется ставить «галочки» одновременно во многих полях. Имеется возможность указывать семантику и на уровне отдельной лексемы. Обязательный «перевод» диалектного слова осуществляется только при наличии соответствующего словаря или примечаний к текстам, сделанных лицом, предоставившим эти тексты. Лингвист, размечающий тексты, не берет на себя ответственность по «переводу» на литературный язык. Тем не менее при разметке диалектной лексемы могут быть введены как словообразовательные, так и близкие по семантике корневые дериваты, чтобы в случае поиска появлялись и незапрашиваемые первоначально диалектные слова с тем же корнем (слева – слово из поисковика, справа – возможные «добавления»):

теперь = *теперича*, *тепере*
вместе = *вместях*
назавтра = *назавтре*
сперва, сначала = *напере*, *перва*

накануне = *конь*
каждый = *кажной*
свекровь = *свекровушка*, *свекровка*, *свекрова*
свадьба = *свайба*
деревня = *деревнюшка*
разделяться = *разделиваться* и т.д.

В Основном корпусе НКРЯ достаточно нажать на любое слово внутри текста, чтобы получить доступ к семантике через выход на размещенные в Интернете онлайн-словари литературного языка. К сожалению, пословного выхода на доступные лексикографические электронные ресурсы у диалектологов пока нет. «Словарь русских народных говоров» (СРНГ) размещен на сайте ИЛИ РАН [16], «Архангельский областной словарь» – на сайте филологического факультета МГУ им. М.В. Ломоносова [17]. Нет постоянной «прописки» и, следовательно, постоянного доступа ко многим диалектным словарям, которые существуют в электронном виде и «гуляют» по Интернету: к Словарю д. Деулино, Большому Донскому словарю, Словарю русских говоров Карелии и многим, многим другим. Но даже когда все эти словари будут находиться в постоянном онлайн-доступе, необходимо проделать огромную работу, чтобы от лексемы в тексте иметь возможность выйти на соответствующую лексему в диалектных словарях и увидеть круг значений.

**5. Банк диалектных текстов в НКРЯ**

«Корпус диалектных текстов» НКРЯ предполагает включение любых диалектных текстов на русском языке, записанных как на территории исконного проживания русского населения (центр европейской части 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.).

Текст в Диалектном подкорпусе может выдаваться в виде «контекста» – при поиске лексемы или конкретной грамматической характеристики пользователь получает предложение с искомым словом + ближайший левый и правый контекст (как это устроено в Основном корпусе). Однако пользователь может по запросу получить целый текст: в устном тексте границы предложений условные, а для анализа той или иной формы или словосочетания нужен очень ши-

рокий дискурсивный контекст. 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). Пока Текст-2 на сайте показываться не будет (хотя информация о нем сохранена в пословной разметке и может быть извлечена). Но там, где тексты первоначально были поданы в орфографизированной записи, Текст-1 и Текст-2 практически совпали.

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. Имеется достаточно большое количество текстов, опубликованных типографским способом и до сих пор не оцифрованных: это образцы говоров из хрестоматий по русской диалектологии, из учебников и пособий по русской диалектологии, пособий по изучению региональной лексики, различных сборников и статей по русской диалектологии. Все эти материалы требуется первоначально оцифровать, и время на их ручной ввод или исправление текстов, полученных путем автоматического распознавания, примерно одинаковое, так как диалектные тексты, как правило, подаются в транскрипции.

В декабре 2016 г. Диалектный корпус пополнился текстами, записанными в Архангельской, Тверской, Тамбовской, Тюменской, Самарской, Волгоградской областях, в Ставрополье, Забайкалье и некоторых других местах. В корпус вошли материалы из следующих книг и хрестоматий: *Кудряшова Р.И.* Донские казачьи говоры Волгоградской области (тексты и задания к ним): учебно-методическое пособие. Волгоград, 2010; *Юмсунова Т.Б.* Язык семейских – старообрядцев Забайкалья. М., 2005; Русская диалектология: учеб. пособие для практических занятий / под ред. Е.А. Нефедовой. М., 1999. 2-е изд. М., 2005. Выставлены не опубликованные ранее экспедиционные материалы: из Самарской обл. (предоставлены Т.Ф. Зибровой и Т.Е. Баженовой; из Тамбовской обл. (предоставлены Т.В. Махрачевой); из Тверской обл. (предоставлены А.И. Рыко); из Ставрополья (предоставлены В.М. Грязновой); из Архангельской обл. (предоставлены И.Б. Качинской); из Саратовской обл. (предоставлены

В.Е. Гольдиным). Материалы из Тверской обл. (говоры Селигера) оказалось возможным дополнить звукорядом, т.е. уже сейчас можно прослушать аудиофайл наряду с просмотром расшифровок. В самом скором будущем появится возможность и видеосопровождения.

Размечены и подготовлены к вывеске материалы из книг и хрестоматий: *Мызников С.А.* Русские говоры Среднего Поволжья. Чувашская республика Марий Эл. СПб., 2005;. *Белякова С.М.* Русская диалектология (лексика): учеб. пособие. Тюмень: Изд-во Тюм. гос. ун-та. 2011; *Иванцова Е.В.* Живая речь русских старожилов Сибири: сб. текстов. Томск, 2007; *Здобнова З.П.* Хрестоматия по русской диалектологии (на материале русских говоров Башкирии). Уфа, 2010; *Русская* речь Коми-Пермяцкого округа: Хрестоматия / сост.: И.И. Бакланова, О.В. Гордеева. А.С. Лобанова, И.А. Подюков, И.И. Русинова. Пермь, 2013; *Долгушев В.Г.* Хрестоматия вятских говоров: Лексика. Тексты. Контрольные задания для студентов: пособие для практических занятий по курсу «Русская диалектология». Киров, 2009. Экспедиционные материалы из Тюменской обл. (подготовленные и предоставленные Е.П. Багировой), из Архангельской, Тамбовской обл., говоры Среднего Урала, русские говоры Азербайджана.

Готовятся к разметке хрестоматии Н.Н. Дурново и Д.Н. Ушакова, Г.Г. Мельниченко, В.И. Трубинского («Русская диалектология: говорит бабушка Марфа, а мы комментируем». СПб., 2004), Р.Ф. Касаткиной (Русские народные говоры. Звучащая хрестоматия: Южнорусское наречие. М., 1999), Л.П. Батыревой (Хрестоматия по диалектологии. Говоры Владимирско-Поволжской группы: записи устной речи и письменные источники. Шуя, 2007) и некот. др.)

И все же на сегодняшний день портфель Корпуса диалектных текстов НКРЯ представляется достаточно «тощим». Мы обращаемся с большой просьбой ко всем диалектологам, держателям текстов, передавать тексты в Диалектный подкорпус. Ведь доступ к диалектным текстам до сих пор значительно 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. В настоящее время 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 большие фонотеки, расшифрованные лишь в малой части.

В благодарность за предоставленные материалы коллектив Диалектного подкорпуса составляет письма в адрес руководителей соответствующих вузов (учреждений) от Института русского языка им. В.В. Виноградова РАН.

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, материала для научной и педагогической работы, демонстрации этнографических, этнокультурных традиций, особенностей русского менталитета.

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

***Литература и электронные ресурсы***

1. *Школьный* диалектологический атлас «Язык русской деревни». – URL: http://gramota.ru/book/village (Институт русского языка им. В.В. Виноградова РАН, (дата обращения: 20.12.2016).

2. *Фонетика* русских диалектов. – URL: http://dialect.philol.msu.ru/index.php (МГУ им. М.В. Ломоносова (дата обращения: 2012.2016).

3. *Диалектная* фонетика. Акустическая база данных по русским говорам. – URL: http://dialect-phon.ruslang.ru/ (Институт Славяноведения РАН) (дата обращения: 20.12.2016).

4. *Информационный* центр «Русская диалектология». – URL: http:// www.ruslang.ru/agens.php?id=rus_dialectology, (Институт Русского языка им. В.В. Виноградова РАН (дата обращения: 20.12.2016).

5. *Электронная* библиотека русских народных говоров. – URL: http://dialekt.rx5.ru/index.html (Казанский (Приволжский) федеральный университет (дата обращения: 20.12.2016).

6. *Лингвогеографическая* система «Диалект». – URL: http://io.udsu.ru/dl/common.logon (Ижевск, Удмуртский госуниверситет (дата обращения: 20.12.2016).

7. *Говор* бассейна Устьи: Корпус севернорусской диалектной речи. – URL: http://www.slavist.de/Pushkino/login.php (Ustja River Basin Corpus Query interface, Р.фон Вальденфельс, Берн, Швейцария и Н.Р. Добрушина и М.А. Даниэль, Высшая школа экономики, Москва (дата обращения: 20.12.2016).

8. *Электронные* базы данных по русским народным говорам. – URL: http://starling.rinet.ru/cgi-bin/main.cgi?root=ruscorpora&encoding=utf-rus (тексты, записанные в деревнях Харовского района Вологодской обл. и Шатурского района Московской обл., С.А. Крылов и А.В. Тер-Аванесова (дата обращения: 20.12.2016).

9. *Региональная* этнолингвистика. – URL: http://www.ethnolex.ru/ (русские говоры Кубани, дата обращения: 2012.2016).

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

11. *Летучий А. Б.* Диалектный корпус: состав и особенности разметки // Национальный корпус русского языка. – Новые результаты и перспективы. – СПб., 2006–2008. С. 114–128.

12. *Качинская И.Б.* Корпус Диалектных Текстов в Национальном корпусе русского языка: состояние и перспективы // Лексический атлас русских народных говоров: материалы и исследования. 2009. – СПб., 2009. –С. 57–68.

13. *Качинская И.Б., Моисеева Е.В.* Диалектный Подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место // Русская устная речь: материалы междунар. науч. конф. «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения». Межвузовское совещание «Проблемы создания и использования диалектных корпусов», Саратов, 15–17 ноября 2010 г. / ред. О.Ю. Крючкова, А.И. Буранова, В.Е. Гольдин, Л.В. Балашова. – Саратов, 2011. – С. 245–255.

14. *Качинская И.Б., Сичинава Д.В.* Корпус диалектных текстов в национальном корпусе русского языка: сегодняшнее состояние и перспективы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог», Бекасово, 4–8 июня 2014 г. – М., 2014. – Вып. 13 (20). – С. 593–600.

15. *Качинская И.Б., Сичинава Д.В.* Диалектный подкорпус сегодня // Тр. Института русского языка им. В.В. Виноградова. – Вып. 6. – М., 2015. – С. 142–162.

16. *Словарь* русских народных говоров. – URL: http://iling.spb.ru/vocabula/srng/srng.html (дата обращения: 20.12.2016).

17. *Архангельский* областной словарь. – URL: http://www.philol.msu.ru/~dialectology/dictionary/ (дата обращения: 20.12.2016).

**ON THE CORPUS OF DIALECTAL TEXTS IN THE RUSSIAN NATIONAL CORPUS**

*Voprosy leksikografii – Russian Journal of Lexicography*, 2017, 11, pp. 71–85.
DOI: 10.17223/22274200/11/5

*Irina B. Kachinskaya,* Lomonosov Moscow State University (Moscow, Russian Federation). E-mail: kacza@yandex.ru

*Dmitry V. Sichinava,* Vinogradov Institute of the Russian Language of the Russian Academy of Sciences (Moscow, Russian Federation). E-mail: mitrius@gmail.com

**Keywords**: Russian dialectology, corpus linguistics, lexical semantics, electronic resources, morphological marking.

The paper deals with the present state of the Corpus of Dialectal Texts within the Russian National Corpus. The Dialectal Corpus is available online at the site http://www.ruscorpora.ru/search-dialect.html. It is searchable since December 2006. As time passed the markup team has changed and so did the tenets of the markup. The new team has developed a new standard of formatting the dialectal texts. According to the latter, texts should be included into the corpus in a phonetic representation, with marked stress, and the user should have the opportunity to work both with fragments and with whole texts. In the paper the main principles of grammatical, semantic and metatext markup of the dialectal texts are described, as well as the guidelines for online search.

The metatext markup consists of three levels: 1) provenance of a text; 2) phonetic markup; 3) genre and topic. A subcorpus can be customized basing on any combination of these sets of parameters. It is possible to select the texts with orthographical rendering and/or with audio recording available.

All the texts within the Dialectal Subcorpus are with resolved morphological ambiguity, with full morphological markup, including the dialectal characteristics. The search, as within the bulk of the RNC, is operated in two regimes: as an exact (sub)string and by lemmata and grams.

The semantic annotation is two-tiered, represented both in the metatext tagging and as a part of the word-by-word annotation. The topic (aboutness) of the text is determined on the metatext level. A separate lexeme can be also tagged semantically. A word is "translated" into the standard Russian language only if the text has a dictionary or notes annexed by the transcriber. It is possible also to add derivates to find other dialectal words with the same root.

In 2016 the Dialectal Subcorpus was updated and now has 300 thousand items.

The Dialectal Subcorpus of the RNC supposes inclusions of every sort of dialectal texts available in Russian, from the historical Russian area (Central European Russia), from the early colonization area (North European Russia) and late colonization area (Siberia, Far East, the Don, the South Volga), as well as the Russian diaspora, mostly Old Believers and Protestants (Latgale, Azerbaijan, Romania, Australia, Canada, the United States and others). The texts are provided by dialectologists who do fieldwork. They can provide transcripts from their personal notes or audio recordings, as well as published texts.

The authors hope that the Dialectal Corpus will soon become a representative collection and will be widely accessed by users.

### *References*

1. Vinogradov Institute of Russian Language, RAS. (n.d.) *Shkol'nyy dialektologicheskiy atlas "Yazyk russkoy derevni"* [School dialectological atlas "Language of the Russian Village"]. [Online] Available from: http://gramota.ru/book/village. (Accessed: 20.12.2016).

2. Lomonosov Moscow State University. (n.d.) *Fonetika russkikh dialektov* [Phonetics of Russian dialects]. [Online] Available from: http://dialect.philol.msu.ru/index.php. (Accessed: 20.12.2016).

3. Institute of Slavic Studies, RAS. (n.d.) *Dialektnaya fonetika. Akusticheskaya baza dannykh po russkim govoram* [Dialectal phonetics. Acoustic database on Russian dialects]. [Online] Available from: http://dialect-phon.ruslang.ru/. (Accessed: 20.12.2016).

4. Vinogradov Institute of Russian Language, RAS. (n.d.) *Informatsionnyy tsentr "Russkaya dialektologiya"* [Information Center "Russian Dialectology"]. [Online] Available from: http://www.ruslang.ru/agens.php?id=rus_dialectology. (Accessed: 20.12.2016).

5. Kazan Federal University. (n.d.) *Elektronnaya biblioteka russkikh narodnykh govorov* [Electronic library of Russian folk dialects]. [Online] Available from: http://dialekt.rx5.ru/index.html. (Accessed: 20.12.2016).

6. Udmurt State University. (n.d.) *Lingvogeograficheskaya sistema "Dialekt"* [Linguistic system "Dialect"]. [Online] Available from: http://io.udsu.ru/dl/common.logon. (Accessed: 20.12.2016).

7. Waldenfels, R.F., Dobrushina, N.R. & Daniel, M.A. (n.d.) *Ustja River Basin Corpus Query interface*. [Online] Available from: http://www.slavist.de/Pushkino/login.php. (Accessed: 20.12.2016).

8. Krylov, S.A. &Ter-Avanesov, A.V. (n.d.) *Elektronnye bazy dannykh po russkim narodnym govoram* [Electronic databases on Russian folk dialects]. [Online] Available from: http://starling.rinet.ru/cgi-bin/main.cgi?root=ruscorpora&encoding=utf-rus. (Accessed: 20.12.2016).

9. Ethnolex.ru. (n.d.) *Regional'naya etnolingvistika* [Regional ethnolinguistics]. [Online] Available from: http://www.ethnolex.ru/. (Accessed: 20.12.2016).

10. Letuchiy, A.B. (2005) Korpus dialektnykh tekstov: zadachi i problemy [Corpus of dialect texts: tasks and problems]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [The National Corpus of the Russian language: 2003–2005. Results and prospects]. Moscow: Indrik.

11. Letuchiy, A.B. (2008) Dialektnyy korpus: sostav i osobennosti razmetki [Dialectical corpus: composition and features of markup]. In: *Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [The National Corpus of the Russian language: 2006–2008. New results and prospects]. St. Petersburg: Nestor-Istoriya.

12. Kachinskaya, I.B. (2009) Korpus Dialektnykh Tekstov v Natsional'nom korpuse russkogo yazyka: sostoyanie i perspektivy [Corpus of Dialect Texts in the National Corpus of the Russian Language: State and Prospects]. In: *Leksicheskiy atlas russkikh narodnykh govorov (Materialy i issledovaniya)* [Lexical Atlas of Russian Folk Dialects (Materials and Research)]. St. Petersburg: Nestor-Istoriya.

13. Kachinskaya, I.B. & Moiseeva, E.V. (2011) [Dialectical Subcorpus of the NCRL. A new filing standard. A new workplace]. *Russkaya ustnaya rech'* [Russian Oral Speech]. Proceedings of the international conference. Saratov. 15–17 November 2010. Saratov: Saratov State University. pp. 245–255. (In Russian).

14. Kachinskaya, I.B. & Sichinava, D.V. (2014) [Corpus of dialect tests in the National Corpus of the Russian language: current state and prospects]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computer linguistics and intelligent technologies]. Based on the proceedings of the annual innternational conference Dialog (2014). Bekasovo. 4–8 June 2014. Vol. 13 (20). Moscow: RSUH. pp. 593–600. (In Russian).

15. Kachinskaya, I.B. & Sichinava, D.V. (2015) Dialektnyy podkorpus segodnya [Dialectic subcorpus today]. *Trudy instituta russkogo yazyka im. V.V. Vinogradova*. 6. pp. 142–162.

16. Iling.spb.ru. (n.d.) *Slovar' russkikh narodnykh govorov* [Dictionary of Russian folk dialects]. [Online] Available from: http://iling.spb.ru/vocabula/srng/srng.html. (Accessed: 20.12.2016).

17. Lomonosov Moscow State University, Faculty of Philology. (1980–2015) *Arkhangel'skiy oblastnoy slovar'* [Arkhangelsk regional dictionary]. [Online] Available from: http://www.philol.msu.ru/~dialectology/dictionary/. (Accessed: 20.12.2016).