КОРПУС ДИАЛЕКТНЫХ ТЕСТОВ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: СЕГОДНЯШНЕЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ[1]

Сичинава Д. В. (mitrius@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Качинская И. Б. (kacza@yandex.ru)

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

Диалектные тексты, особенно в транскрипционной записи, часто оказываются малодоступными даже для специалистов. Уже само название Диалектного подкорпуса НКРЯ — «Корпус диалектных текстов» — указывает на то, что Пользователю предоставляется возможность работать с цельными текстами, записанными в полевых условиях в разное время. В докладе рассказывается о подготовке текстов к размещению на сайте НКРЯ в программе «Рабочее место диалектолога», с помощью которой осуществляется и разметка на уровне грамматики, в том числе с указанием диалектных особенностей, и метаразметка: указывается «паспорт» текста (место, время, автор записи, сведения об информантах и проч.), отмечаются его фонетические особенности, жанровая и тематическая отнесенность.

Ключевые слова: Корпусная лингвистика. Русская диалектология. Национальный корпус русского языка. Диалектный подкорпус

[1] Работа над Диалектным подкорпусом НКРЯ поддержана грантом РГНФ № 14-04-12012, проект «Корпус диалектных текстов Национального корпуса русского языка: пополнение и разметка», рук. Д. В. Сичинава.

THE DIALECTAL SUBCORPUS WITHIN THE RUSSIAN NATIONAL CORPUS: TODAY AND TOMORROW

Sitchinava D. V. (mitrius@gmail.com)
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

Kachinskaya I. B. (kacza@yandex.ru)
Lomonosov Moscow State University, Moscow, Russia

The main results of the project aimed at developing the dialectal subcorpus of the RNC were the creation of a pilot corpus and the change of the markup principles encompassing many dialectoloigical parameters. A working place program was created and many texts were marked up using the new technology. The present goal of our team is a considerable increase of the corpus, its representativeness and the depth of linguistic processing. The dialectal texts available for search in the RNC (www.ruscorpora.ru/search-dialect.html) will be considerably updated, with the overall corpus size reaching 1 mln tokens. The texts, mainly unpublished or published in rather obscure editions, are to be made available for a wider circle of dialectologists. Some texts are to be accompanied with video and audio. Alongside with word-by-word grammarical markup with resolved homonymy, the texts are to be tagged extensively on the metalevel (data of creation, dialect, overall phonetical properties and others). The accumulation of dialectal texts will be continued, the dialectologists who had collected valuable texts are invited to share their results with the professional community.

Kew words: Russian dialectology, corpus linguistics, Russian National Corpus

1. Существующие наработки в области диалектных корпусов

«Корпус диалектных текстов» входит в состав Национального Корпуса Русского языка (НКРЯ). Этот корпус сопоставим с такими известными национальными корпусами, как Британский и Чешский. За рубежом существуют и корпуса, включающие диалектные тексты — например, корпус, созданный в Китае (в рамках «Программы 863») или в странах Скандинавии (корпус http://www.tekstlab.uio.no/nota/scandiasyn/); много занимаются изучением народных говоров в Польше (http://www.dialektologia.uw.edu.pl/index.php); недавно появился корпус грузинских диалектов (http://mygeorgia.ge/gdc/).

В России интерес к прикладной лингвистике, созданию словарей, в том числе диалектных, сопровождается также и большой работой по созданию Диалектных корпусов, которые далеко не всегда имеют локальную

ограниченность. Так, создан сайт «Школьный диалектологический атлас "Язык русской деревни"»: http://gramota.ru/book/village. Имеются значительные по объему корпуса диалектных текстов в Казани: «Электронная библиотека русских народных говоров», Казанский (Приволжский) федеральный ун-т, где собраны материалы многих экспедиций в различных говорах Европейской части России (http://dialekt.rx5.ru/index.html); в Ижевске — Лингвогеографическая система «Диалект», Удмуртский ун-т (http://lgw2.udsu.ru:9001/). В интернете появились тексты из Шатурского р-на Московской обл. и Харовского р-на Вологодской обл. в рамках проекта «Электронные базы данных по русским народным говорам» (авторы — С. А. Крылов и А. В. Тер-Аванесова — http://starling.rinet.ru/cgi-bin/main.cgi?root=ruscorpora&encoding=utf-rus). Открылся новый сайт «Региональная этнолингвистика» с материалами по кубанским говорам (http://www.ethnolex.ru/). Во многих вузах продолжается работа по созданию и совершенствованию корпусов (пока без доступа в Интернете). По материалам трех русских говоров (двух южных и одного северного) 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). Материал более чем из ста говоров Архангельской области содержится в корпусе «Электронная картотека „Архангельского областного словаря“» (МГУ имени М. В. Ломоносова), общий общем которого приближается к 2-м млн «карточек» — как видно из названия, этот корпус имеет жесткую лексикографическую направленность. Вышло 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. Ведется активная работа по созданию корпусов по русским народным говорам в Томске, Тюмени, Челябинске, Смоленске и других научных центрах, организованных на базе университетов (см., напр., [Русская устная речь, 2011; Юрина, 2011]).

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

2. Концепция корпуса

«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 Латгалии, Азербайджана, Румынии, Австралии, Канады, Америки). Туда войдут полевые записи, аудио- и видеорасшифровки, тексты из хрестоматий, малодоступных и малотиражных сборников и изданий. Мы надеемся, что со временем этот

корпус станет репрезентативным собранием диалектных текстов и будет одним из самых посещаемых и востребованных пользователями.

Работа над Диалектным корпусом уже была поддержана грантами РГНФ: в 2006–2008 гг. (№ 06-04-13818в: «Создание корпуса диалектных и фольклорных текстов на русском языке», рук. В. М. Живов) и в 2009–2010 (№ 09-04-12159в: «Корпус диалектных текстов Национального корпуса русского языка: Грамматическая, фонетическая и метатекстовая разметка. Новый стандарт подачи», рук. В. М. Живов). Результатом первого (пилотного) проекта было создание Диалектного подкорпуса в составе НКРЯ (см. [Летучий 2005; 2008]), результатом второго проекта 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 ударениями в фонетической транскрипции двух видов: «начальной» и «облегченной», унифицированной; была значительно усовершенствована грамматическая разметка; пополнился банк диалектных текстов; часть текстов публикуется в новом формате [Качинская, 2009; 2011].

Программа, в которой непосредственно производится разметка и метаразметка диалектных текстов, — среда «Рабочее место диалектолога» (автор Т. А. Архангельский), — находится в свободном доступе.

Диалектные тексты, особенно в транскрипционной записи, часто оказываются малодоступными даже для специалистов. В ближайшее время наша задача состоит в резком увеличении количества текстов, включенных в Диалектный корпус НКРЯ (www.ruscorpora.ru/search-dialect.html) и репрезентативности географии записи. Фундаментальность проекта связана не только с количеством текстов, но и со степенью их лингвистической обработки: все тексты будут представлены с так наз. «снятой омонимией»: разметка осуществляется как на уровне грамматики (в том числе с указанием диалектных особенностей на уровне слова), так и на уровне метаразметки с указанием особенностей жанра, тематики и фонетических особенностей (в самых общих чертах). Лингвисты и все любители народного слова получат широкий доступ к цельным диалектным текстам: некоторые из них хотя и публиковались типографским способом, но обычно малыми тиражами и в малодоступных изданиях; многие тексты ранее нигде не публиковались. Некоторые тексты, записанные в последние годы на цифровые носители, предполагается сопровождать аудио- и видеоматериалами.

3. Пополнение и развитие корпуса на текущем этапе

Создание диалектных корпусов — дело во многом новое. Сбор диалектных текстов для включения их в корпус на сайте НКРЯ не должен препятствовать созданию диалектных корпусов, так сказать, «местного масштаба». С одной стороны, мы хотели бы задать некоторый стандарт подачи диалектного текста — речь идет прежде всего о текстах, которые будут специально расшифровываться для корпуса. С другой стороны, нельзя игнорировать уже имеющиеся

образцы записи народных говоров, которые оказались далеки от нашего «стандарта». Главное отличие Диалектного корпуса НКРЯ от других русских диалектных корпусов видится нам в установке на **сплошную грамматическую разметку** текстов, что соответствует общей стратегии всего Национального корпуса в целом, тогда как региональные корпуса, скорее всего, в основу разметки будут класть приципы семантики.

Работа по пополнению корпуса является достаточно сложной. Уже с самого начала эту работу приходится разделить на 2 составляющие. Первая часть — **создание банка диалектных текстов**. Тексты предоставляются диалектологами, ведущими полевую работу. Это следующие типы текстов:

1) Записи из полевых тетрадей или аудиорасшифровок, уже введенные в компьютер.
2) Опубликованные тексты, предоставленные в компьютерном варианте.
3) Тексты, опубликованные типографским способом и до сих пор **не** введенные в компьютер: 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 ввести в компьютер, и время на их ручной ввод или исправление текстов, полученных путем автоматического распознавания, примерно одинаковое, т. к. диалектные тексты, как правило, подаются в транскрипции с использование диакритик.

Мы также надеемся, что для Диалектного корпуса НКРЯ будут специально делаться также

4) расшифровки аудио- и видеофайлов в заданном нами стандарте фонетической транскрипции (на уровне Текста-1 или Текста-2) в шрифтах юникода.

Вторая часть работы, связанная с лингвистической обработкой текстов, тоже достаточно трудоемка.

1) Сначала требуется **подготовить** текст для его обработки в среде «Рабочее место диалектолога» (РМ), для чего необходимо отделить текст информанта от текста собирателя (квадратными скобками); проставить пробелы перед дефисами — например, перед постчастицами и во всех других случаях, если словоформу необходимо разбирать как 2 слова, а не как одно (*г дому-ту, чему-то, дедушко- домовеюшко*); перевести поданную держателем текстов транскрипцию (иногда изготовленную в самостийных шрифтах или с использованием сразу нескольких шрифтов) в единый шрифт юникода; перевести текст в формат txt в кодировке UTF-8.
2) При открытии уже подготовленного текста в РМ автоматически срабатывают специально созданные **детранскрипторы**: детранскриптор-1 переводит Текст-1 в Текст-2 — первоначальную транскрипцию в «облегченную», унифицированную, детранскриптор-2 переводит Текст-2 в Текст-3 — орфографизированный вариант (= орфографический «подстрочник»), необходимый для работы стандартного

грамматического анализатора. Текст-3 необходимо вручную довести до уровня орфографии, хотя и здесь программа предусматривает некоторую помощь размечающему текст диалектологу: при нажатии кнопки ОРФО (на центральной горизонтальной панели) в Тексте-3 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.

3) После обработки Текста-3 осуществляется **грамматическая разметка** текста. Кнопка XML, расположенная на центральной горизонтальной панели РМ на уровне ТЕКСТ, осуществляет прогон грамматического анализатора по всему тексту (работает по Тексту-3), 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. Т.е. поверх стандартной грамматической разметки в диалектных текстах предусмотрена возможность отмечать диалектные грамматические особенности лексемы. Для этого в РМ внедрены грамматические таблицы по каждой из 5 изменяемых частей речи (глаголы, существительные, прилагательные, местоимения, числительные).

Если в предыдущем проекте (2008–2009 гг.) нами использовались таблицы с реальными диалектными аффиксами, формами и проч., то к сегодняшнему дню мы отказались от навязчивой конкретики, так как при работе с текстами выяснялось, что списки аффиксов всегда оказывались неполными. В связи с этим большую часть текстов, уже подготовленных для размещения на сайте НКРЯ в рамках предыдущего проекта, пришлось переделывать.

Для каждой изменяемой части речи предусмотрена возможность указывать диалектные особенности на следующих уровнях:

глагол: основа — флексия — суффикс — форма — вид — переходность — возвратность — время. Так, например, диалектные особенности инфинитива отмечаются либо указанием на **диалектный суффикс** (*пекчи*), либо **диал. основу** (*трать, жмать*), **диал. форму** (*идтить*). То же с императивом, где диал. особенности могут проявляться либо на уровне **суффикса** (*посодь*), либо на уровне **формы** (*доедь, ехай*). Иногда лучше отметить сразу несколько возможностей — диал.суффикс и диал. основу (*посодь*).

Особенности спряжения можно указывать следующим образом, например:
общее спряж. (*любют*): *любить* … наст.вр. 3 л. мн. ч. + **диал. флексия**
3-е спряж. (*они гулят*): *гулять*… наст.вр. 3 л. мн. ч. + **диал.флексия**

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

Кнопка ФОРМА зарезервирована для случаев, когда в ЛЯ нет соответствий или трудно/невозможно разделить основу и флексию: типа *jo* или *jy* как формы местоимения *её = она,* 3 л.ж.р. Вин. ед., или *ихной, ихний, ихой* (= *их*), *тэй / тый парень* (= *тот*), *оне, оны* (= *они*), сравн. ст. прилаг. / наречий — *первеющий* и проч.

4) Помимо грамматической разметки, для каждого диалектного текста осуществляется **метаразметка**, которая содержит три уровня:

1 — Адрес-сопровождение
2 — Фонетическая
3 — Диалектная текстовая

Пока что 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 неполное оканье или диссимилятивное аканье);
(3) в области консонантизма: <г> взрывной или фрикативный и
(4) цоканье.

В банке текстов, переданных для размещения 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.

Если текст, поданный в транскрипции (или в орфографии), будет сопровожден **фонетическим** комментарием, предоставленным держателями текстов, эта информация будет доступна пользователю, т. е. будет активна кнопка «Комментарии к тексту». Это касается и любых других комментариев, связанных с **грамматикой** или **семантикой** предоставленных текстов, и, возможно, даже с какой-то **экстралингвистической** информацией (этнографической, этнокультурной). Можно сопровождать тексты фотоматериалами. Предполагается включать также **аудио- и видеосопровождение**.

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

ЖАНР (ТИП) ТЕКСТА делится на 4 категории:

- устные нефольклорные тексты
- письменные нефольклорные тексты
- устные фольклорные тексты
- письменные фольклорные тексты

Пока предпочтение с точки зрения включения в корпус отдается устным нефольклорным текстам, хотя и там могут содержаться элементы фольклора, жанры которых отмечаются не в пределах встречаемости, а в пределах всего

текста в целом (в устном рассказе естественно могут оказаться и колыбельные песни, и частушки, и пословицы, поговорки, загадки, и проч.).

В то же время в банке диалектных текстов уже есть как письменные фольклорные тексты (напр., «песенники» или заговоры, записанные самими носителями), так и письменные нефольклорные (письма, мемуары, дневники и проч.).

Распределение текстов по **тематике** и **семантике** осуществляется, но требует доработки.

Мы пытались ввести заинтересованного читателя (а также зрителя и слушателя) в свою «лабораторию», обратить внимание на некоторые, порой неожиданно возникающие сложности работы с текстами; показать, как сейчас обрабатываются диалектные тексты для НКРЯ и как они будут выглядеть на сайте в ближайшее время.

Свободное предоставление в Интернете текстов русских народных говоров, а также их грамматическая, семантическая и метатекстовая характеристика позволит специалистам-диалектологам, другими лингвистам и нелингвистам, филологам, историкам, культурологам, этнографам — и 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 и проч. Некоторые тексты предполагается сопровождать звуко- и видеорядом (в случае, когда тексты явились расшифровками аудио- и видеозаписей). В последующем планируется создание серии интерактивных карт с указанием точки на карте, соответствующей данному пункту с демонстрацией запрашиваемого явления на карте в масштабе области / Европейской части РФ / всей России.

Литература

1. *Качинская И. Б.* (2009), Корпус диалектных текстов в Национальном корпусе русского языка: состояние и перспективы, Лексический атлас русских народных говоров (Материалы и исследования), СПб., с. 57–68. (http://www.philol.msu.ru/~ruslang/pdfs/kachinskaya.i.b/19.pdf от 01.03.2014)
2. *Качинская И. Б.* (2011), Диалектный подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место, Русская устная речь. Материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения». Межвузовское совещание «Проблемы создания и использования диалектных корпусов», СГУ, Саратов, с. 245–255.
3. *Летучий А. Б.* (2005), Корпус диалектных текстов: задачи и проблемы (http://ruscorpora.ru/sbornik2005/13letuchy.pdf), Национальный корпус русского языка: 2003–2005, Индрик, Москва. С. 215–232.
4. *Летучий А. Б.* (2009) Диалектный корпус: состав и особенности разметки (http://ruscorpora.ru/sbornik2008/06.pdf), Национальный корпус русского

языка: 2006–2008. Новые результаты и перспективы, Нестор-История, СПб., С. 114–128.
5. *Русская устная речь* (2011), Материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения». Межвузовское совещание «Проблемы создания и использования диалектных корпусов», СГУ, Саратов.
6. *Юрина Е. А.* (2011), Томский диалектный корпус: в начале пути, Вестник Томского гос. ун-та. Филология, № 2, С. 58–63 (http://cyberleninka.ru/article/n/tomskiy-dialektnyy-korpus-v-nachale-puti от 01.03.2014).

References

1. *Jurina E. A.* (2011), The first steps of the Tomsk Dialectal Corpus [Tomskij dialektnyj korpus: v nachale puti], Vestnik Tomskogo gosudarstvennogo universiteta. Filologija, 2, p. 58–63, access at: http://cyberleninka.ru/article/n/tomskiy-dialektnyy-korpus-v-nachale-puti, access date 01.03.2014.
2. *Kachinskaja I. B.* (2009), The corpus of dialectal texts within the Russian National Corpus: current state and plans [Korpus dialektnyh tekstov v Natsional'nom korpuse russkogo jazyka: sostojanie i perspektivy], Studies concerning the Lexical altas of Russian dialects [Leksicheskij atlas russkih narodnyh govorov (Materialy i issledovajia), St. Petersburg, p. 57–68, access at http://www.philol.msu.ru/~ruslang/pdfs/kachinskaya.i.b/19.pdf, access date 01.03.2014
3. *Kachinskaja I. B.* (2011), The dialectal subsorpus of the RNC. The new format and linguist's GUI [Dialektnyj podkorpus NKRJA. Novyj standart podachi. Novoe rabochee mesto], Proceedings of the Conference: Spoken Russian. Dialectal and Colloquial Speech Cultures. Workshop: Building and usage of Dialectal Corpora [Russkaja ustnaja rech'. Materialy mezhdunarodnoj nauchnoj konferencii "Barannikovskie chtenija. Ustnaja rech': russkaja dialektnaja i razgovorno-prostorechnaja kul'tura obschenija". Mezhvuzovskoe soveshchanie "Problemy sozdanija i ispol'zovanija dialektnyh korpusov"], Saratov University, Saratov, p. 245–255.
4. *Letuchij A. B.* (2005), Corpus of dialectal texts: tasks and problems [Korpus dialektnyh tekstov: zadachi i problemy], Russian National Corpus 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], Indrik, Moscow, p. 215–232.
5. *Letuchij A. B.* (2009), The dialectal corpus: architecture and tagging properties [Dialektnyj korpus: sostav i osobennosti razmetki], Russian National Corpus 2006–2008: New results and trends [Natsional'nyj korpus russkogo jazyka: 2006–2008: Novye rezul'taty i perspektivy], Nestor-Istorija, Saint-Petersburg, p. 114–128.
6. *Spoken Russian* (2011), Proceedings of the Conference: Spoken Russian. Dialectal and Colloquial Speech Cultures. Workshop: Building and usage of Dialectal Corpora [Russkaja ustnaja rech'. Materialy mezhdunarodnoj nauchnoj konferencii "Barannikovskie chtenija. Ustnaja rech': russkaja dialektnaja i razgovorno-prostorechnaja kul'tura obschenija". Mezhvuzovskoe soveshchanie "Problemy sozdanija i ispol'zovanija dialektnyh korpusov"], Saratov University, Saratov.