РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

ЛЕКСИЧЕСКИЙ АТЛАС РУССКИХ НАРОДНЫХ ГОВОРОВ

(Материалы и исследования)
2009

Санкт-Петербург
«Наука»
2009

И.Б. Качинская

КОРПУС ДИАЛЕКТНЫХ ТЕКСТОВ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: СОСТОЯНИЕ И ПЕРСПЕКТИВЫ*

Корпусная лингвистика развивается в настоящее время мощными темпами. Крупнейшим проектом является общедоступный Национальный корпус русского языка (НКРЯ) www.ruscorpora.ru: «На этом сайте помещен корпус современного русского языка объемом более 140 млн слов. Корпус русского языка – это информационно-справочная система, основанная на собрании русских текстов в электронной форме… Корпус предназначен для всех, кто интересуется самыми разными вопросами, связанными с русским языком: профессиональных лингвистов, преподавателей языка, школьников и студентов, иностранцев, изучающих русский язык»[1].

При создании НКРЯ был учтен опыт работы с Корпусами многих коллективов, в том числе Машинного фонда Института русского языка им. В.В. Виноградова РАН, Лаборатории компьютерной лингвистики Института проблем передачи информации РАН, Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ им.

* Работа поддержана грантом РГНФ № 09-04-12159в («Корпус диалектных текстов Национального корпуса русского языка: грамматическая, фонетическая и метатекстовая разметка. Новый стандарт подачи»).

[1] Все цитаты приводятся по: www.ruscorpora.ru.

М. В. Ломоносова. На сайте НКРЯ имеются отсылки на другие общедоступные корпуса, например: «Упсальский корпус», «Тюбингенский корпус», «Машинный фонд русского языка», «Национальный корпус русского литературного языка (С.-Петербург)», «Регенсбургский диахронический корпус русского языка (древнерусские тексты)», «Рукописные памятники Древней Руси: берестяные грамоты, летописи, рукописная книга» и др.

«Существенной частью поискового аппарата Корпуса является так называемая метаразметка (или метаописание) текстов, входящих в него. Под метаразметкой понимается приписывание тексту атрибутов, характеризующих обстоятельства его создания, автора, тематику, жанровые особенности и др. Метаразметка необходима прежде всего для того, чтобы исследователь, пользующийся Корпусом, мог составлять по своему желанию произвольные выборки текстов с заданными внешними параметрами: например, тексты мемуарного характера, тексты, написанные мужчинами, тексты, написанные авторами, родившимися между 1940 и 1960 гг., тексты автобиографий, тексты проповедей, тексты романов и повестей, и т. д., и т. п.».

Национальный Корпус делится над подкорпусы: Основной, Обучающий, Синтаксический, Параллельный (содержащий переводы с иностранных языков на русский и наоборот), Акцентологический, Поэтический, Устной речи, Диалектный.

Объем Основного корпуса к концу 2008 г. достиг 40 млн слов по текстам первой половины XX в., по текстам XVIII в. приблизился к 3 млн слов. В Корпусе Устной Речи представлено около 8 млн словоупотреблений.

«Корпус устной речи включает в себя расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов. Использована русская стандартная орфография (при этом приводятся наиболее частотные и общепринятые стяженные формы). Возможен лексический, морфологический и семантический поиск, а также формирование пользовательских подкорпусов, в том числе и по социологическим параметрам. Включены тексты самых разных жанров и типов, разного происхождения с точки зрения географии (Москва, Санкт-Петербург, Саратов, Ульяновск, Таганрог, Екатеринбург, Норильск, Воронеж, Новосибирск и мн. др.). Хронологический охват корпуса 1930-2000-е гг.».

Информация об объеме каждого подкорпуса регулярно обновляется на сайте. К моменту публикации этого сборника объем текстов во всех подкорпусах, несомненно, будет значительно увеличен.

В Пилотном проекте Диалектного корпуса, созданном под руководством А. Летучего, каждый представленный текст сопровождается «паспортом», например:

Название *Семья*

Дата создания *2004*
Диалектолог (запись) *И.И. Исаев и др.*
Регион записи *Вологда*
Место записи *Деревня Борок-1*
Сфера функционирования *бытовая*
Тема текста *семья*
Стиль *диалектный*
Предложений *72*
Словоформ *448*
Возраст аудитории *н-возраст*
Уровень образования аудитории *н-уровень*
Размер аудитории *личная*
Источник *Институт русского языка РАН*
Подкорпус *ПК диалектных текстов*
Снятие омонимии *manual*[2].

На конец 2008 г. в этом подкорпусе было представлено ок. 144 тыс. словоупотреблений в 122 текстах, записанных в Архангельской, Брянской, Владимирской, Волгоградской, Вологодской, Воронежской, Ивановской, Калужской, Кировской, Костромской, Курской, Новгородской, Псковской, Рязанской, Саратовской, Смоленской, Тамбовской, Тверской, Тульской областях, а также в Забайкалье и Карелии.

Тексты поданы в орфографизированном виде, отмечены особые диалектные формы (чаще всего грамматические):

нестандартная флексия (*девчонкими*)
другая флексия (*соседы*)
нестандартная форма -ся (*поднялася/-си*)
наличие/отсутствие -ся (*гоститься* 'гостить')
склоняемая частица (*парень-от*)
стяжённая форма прил. (*хорошу*)
формы числа (*польта*)
отличие в роде (*всю лето*)
наличие префикса (*взамуж* 'замуж').

В настоящее время ведется работа над новой редакцией диалектного подкорпуса НКРЯ. Реструктуризация затронет фонетический, географический, грамматический, семантический уровни.

***

1. ФОНЕТИЧЕСКИЙ АСПЕКТ. Проблема возникает уже с принципов подачи материала: если мы публикуем на сайте диалектные тексты, стоит ли сохранять транскрипцию (а если стоит – то какого уровня?

[2] Т.е. омонимы снимались вручную, не автоматически.

насколько «сложную»? «облегченную»?) или вовсе отказаться от попытки передачи фонетического звучания?

Не хотелось бы терять фонетическую информацию – слишком много диалектологами сделано в этом направлении. В новой редакции решено

а) публиковать на сайте текст в том виде, в каком он предоставлен изначально (насколько это позволяют технические возможности), при этом текст будет НЕ В ГРАФИЧЕСКОМ режиме, т.е. НЕ в формате *.pdf. Это Текст-1.

б) Возможно (этот вопрос еще обсуждается) будет существовать ТЕКСТ-2, представленный в «облегченной» транскрипции (ориентированной на принципы русской графики, но с обязательным сохранением ударения) – для удобства цитирования. Например, слово *кон'* будет выглядеть как *конь* - тогда автоматически приходится отказываться от знака **Ь** как обозначающего гласный звук переднего ряда.

Вопросов возникает много: что сохранять в этом «облегченном» (для визуального восприятия и цитирования) варианте? Отказываться от подачи «аканья», т.е. приводить тексты к «оканью» (как это выглядит в Корпусе сейчас) – или сохранять редукцию? Унифицировать ли в одной графеме *г* взрывной и фрикативный? Убирать цоканье или сохранять его (*цяшка* или *чашка*)? Глухой долгий шипящий твердый объединять с мягким (менять *шшуку* на *щуку*)? И что прописывать после мягких *ч, щ, ц* (где они мягкие) – *я, ю* или *а, у*? А там, где они твердые? Заменять ли /ф/ в произношении [хв] (писать *фартук* вместо *хвартук*)? А как быть с изменением групп согласных? Упрощениями, ассимиляцией, диссимиляцией и проч.? *Нощкя, ноцька, куриса, хто, дилектър, амман* (= обман)*, аддать* (отдать)*, бальнёй, поес* (= поезд) *што, штё, цто, цё, чё, ишшо, йешчо* и мн. мн. др. – надо ли все это унифицировать, приводить к орфографической записи? А что делать там, где уже произошла лексикализация или грамматикализация фонетических явлений? И как это понять (произошла она или нет) по одному-двум текстам лингвисту, не знакомому с особенностями конкретного говора?

Мы думаем, что Текст-2 должен, с одной стороны, максимально учитывать особенности транскрипции, с другой – опираться на общепонятную графическую систему русского языка. Можно ориентироваться на достаточное количество публикаций диалектных текстов или на представление материала в диалектных словарях, рассчитанных как на специалистов, так и на неспециалистов.

Может быть, стоит ввести и некий Текст-3, «подстрочник», содержащий грамматический и фонетический аналоги диалектного текста (аналог грамматическим и орфографическим нормам литературного языка; семантический перевод будет даваться отдельно). Тогда текст может выглядеть так:

Текст-1 (исходный. Его подача обязательна).

сев. *А з'орн**о** остай**о**цца на гумн'**е**. Йов**о** обм'ет**а**йут*[3] .

юж. *У гал**о**днъй*[4] *кум'**е** адн**о** нъ ум'**е**.*

Текст-2 (стандартизированная подача в Корпусе, вызывается по требованию пользователя)

сев. *А зёрн**о** остаёцца на гумн**е**. Йов**о** обмет**а**ют.*

юж. *У гал**о**днъй кум**е** адн**о** нъ ум**е**.*

Текст-3 («подстрочник», вызывается по требованию пользователя)

сев. *А зерн**о** остаётся на гумн**е**. Ег**о** обмет**а**ют.*

юж. *У гол**о**дной кум**ы** одн**о** на ум**е**.*

в) Если затранскрибированные тексты решено подавать в имеющемся виде, то необходимо разработать стандарт для новых расшифровок, которые будут (надеемся!) делаться специально для Корпуса диалектных текстов НКРЯ. Транскрипция этих записей должна, с одной стороны,

– достаточно удовлетворять лингвистов,

– быть доступной для чтения широким пользователем-нелингвистом.

г) В рамках существующего проекта хотелось бы учесть огромную работу, проделанную диалектологами по исследованию диалектной фонетики. Представляется возможным осуществлять поиск в текстах не только по грамматическим, но и по фонетическим критериям (разумеется, если предложенный текст фиксирует фонетические особенности говора). Например, в области ударного вокализма – это позиционные изменения гласных /а/, /е/, /старого ятя/ после мягких согласных; в области безударного вокализма – фиксация типов оканья/аканья (без определения конкретных типов диссимилятивного или диссимилятивно-ассимилятивного яканья по отдельно предоставленному тексту. Но если в сопровождающем текст **фонетическом** комментарии будет содержаться информация о конкретном типе диссимилятивного яканья – эта информация вполне может быть предоставлена пользователю).

д) Считаем целесообразным со временем дать возможность пользователю сравнить имеющиеся в Корпусе конкретные диалектные тексты с информацией, содержащейся в ДАРЯ (Диалектном атласе русского языка). Хотелось бы сделать более доступной информацию, над которой трудилось целое поколение диалектологов по всей России. Пока у нас нет технической возможности предоставить собственно карты – надеемся, что когда-нибудь Отдел диалектологии Института русского языка им. В.В. Виноградова (ИРЯ

---

[3] В данной публикации ударения проставлены п/ж шрифтом, но в НКРЯ ударения проставляются стандартным знаком.

[4] И в Тексте-1, и в Тексте-2 в этом случае имеется в виду г-фрикативное. В НКРЯ проблема диакритик (ударений», «крышек») и частотных знаков вроде г-фрикативного или у-неслогового решена в рамках уникода.

РАН) выставит все карты ДАРЯ на своем сайте хотя бы в формате pdf – и тогда мы с удовольствием поставим ссылку на эти карты. Мы думаем над проблемой, как «перевести» лингвогеографические характеристики в характеристики «описательные». Например, по запросу «позиционное изменение /а/ под ударением между мягкими» появляются предложения из тех говоров, где /а/ звучит как [е] (по конкретным материалам, имеющимся в Корпусе), и, кроме того, пользователь может увидеть, что по этому запросу есть в ДАРЯ – не в виде карт, а в виде перечисления районов и областей (*зона произнесения* [е] *на месте /а/ под ударением между мягкими согласными*). Если у нас еще **нет** текстов из какого-то района или тексты представлены не в транскрипции, а в орфографии – мы тоже можем предложить пользователю сведения из ДАРЯ (*зона полного оканья; зона умеренного яканья; зона г-фрикативного*).

Предоставления сведений по ДАРЯ как **фона** для **конкретных** материалов, – это, конечно, вполне самостоятельная работа, и делаться она будет не в первую очередь.

К задачам **не сегодняшнего, но завтрашнего дня** можно отнести видео- и аудиосопровождение текстов;

2. ГЕОГРАФИЧЕСКИЙ АСПЕКТ. Корпус диалектных текстов НКРЯ предполагает включение **любых** диалектных текстов на русском языке, записанных

а) на территории исконного проживания русского населения (Европейская часть России),

б) сопредельных стран (русские говоры Латгалии, Литвы, Эстонии и проч.),

в) раннего заселения (Русский Север),

г) позднего заселения (Сибирь, Дальний Восток),

д) миграций (говоры старообрядцев Румынии, Австралии, Канады, США и проч.).

К несколько более отдаленным по времени задачам относится создание интерактивных карт

– с указанием точки на карте, соответствующей данному пункту;

– с демонстрацией запрашиваемого явления на карте в масштабе

а) области

б) Европейской части РФ

в) всей России

3. ГРАММАТИЧЕСКИЙ АСПЕКТ.

В новой редакции Диалектного корпуса создаются поля для более подробной грамматической метаразметки (с учетом пользовательских запросов), значительно расширяются возможности точечного грамматическо-

го поиска: не просто искать «диалектное окончание» (не совпадающее с литературным), но, например,

а) для существительных I скл. отмечать особый Родительный (*у сестре*), особый Дательный / Предложный в системах, в которых Родительный = Дательному = Предложному (*у Москвы, к Москвы, в Москвы*);

б) ориентацию существительных III склонения на I (*на пече, ночей, ночьей, ночуй*);

в) переход глагола в иной (продуктивный) класс (*продавают, ездиют, полоскает, трать, жмал*) или сохранение старой основы (*мяучит, лачит, полощет*),

г) выравнивание парадигмы в основе настоящего времени (*любю, носю, ездю, пекот*),

д) общее спряжение (*ходют, любют*)

е) III спряжение / тенденция к III спряжению (*играш, играт, играм*) – и т.д.

Одновременно создается рабочее место лингвиста, размечающего тексты, – дружественный интерфейс, чтобы по заранее составленным таблицам можно было делать отметки в всплывающих подсказках-вопросах на всех уровнях разметки и метаразметки (как это во многом уже сделано для пользователя).

4. СИНТАКСИЧЕСКИЙ АСПЕКТ.

Для метатекстовой синтаксической разметки, возможно, будет удобно пронумеровать предложения в тексте и некоторым из предложений дать синтаксическую характеристику. Указывать падежное управление для предлогов в тех случаях, когда это управление не совпадает с литературным или имеет более широкую сферу распространения; отмечать особые синтаксические конструкции.

5. СЕМАНТИЧЕСКИЙ АСПЕКТ.

Ориентируясь на тематические Вопросники, составленные диалектологами для ЛАРНГ, значительно расширить тематическую метатекстовую разметку, а также предоставить пользователю возможность просматривать **словник** встретившихся в текстах лексем, организованный

1) в прямом алфавитном порядке
2) обратном алфавитном порядке
3) с индексом частотности.

При этом пользователь, создав свой корпус (по команде *создать подкорпус*) может ограничить словник интересующими его географическими границами наречий / областей / районов.

Предоставить пользователю словарь-конкорданс диалектизмов (с «рабочими» толкованиями, вытекающими из значения слова в зафиксированных контекстах).

В марте 2009 г. лингвистам-диалектологам и во многие диалектологические центры было разослано письмо с призывом присылать имеющиеся материалы для обработки и размещения их на сайте Диалектного подкорпуса НКРЯ. Пользуясь новой возможностью, мы хотели бы повторить свою просьбу:

Уважаемые коллеги!

Коллектив **Национального корпуса русского языка** (НКРЯ, www.ruscorpora.ru) предлагает вам принять участие в пополнении **Корпуса диалектных текстов** (http://www.ruscorpora.ru/search-dialect.html)

«Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т. п. …Национальный корпус предназначен в первую очередь для обеспечения научных исследований лексики и грамматики языка, а также тонких, но непрерывных процессов языковых изменений, происходящих в языке на протяжении сравнительно небольших периодов — от одного до двух столетий. Другая задача корпуса — предоставление всевозможных справок, относящихся к указанным областям (лексика, грамматика, акцентология, история языка). Современные компьютерные технологии многократно упрощают и ускоряют процедуры лингвистической обработки больших массивов текстов. Раньше исследователь мог лишь просматривать тексты и вручную выписывать из них нужные примеры; эта предварительная (но абсолютно неизбежная) деятельность была очень трудоемкой и не позволяла обрабатывать большие массивы материала. Теперь ограничений на объем анализируемого материала и скорость поиска информации в нем по существу нет, а это означает, что в распоряжении исследователя оказываются колоссальные массивы текстов самого разного типа…»

На сайте НКРЯ ваши материалы (со ссылкой на предоставившее их лицо) станут доступны многим специалистам и любителям русского языка. Публикация материалов на сайте НКРЯ не помешает вам опубликовать эти же материалы на собственных сайтах или в печати – в таком случае мы добавим ссылку на публикации.

1. Тексты могут представлять собой расшифровку аудио- или видеозаписей, а также извлечения из полевых экспедиционных тетрадей.

2. Тексты принимаются в транскрипции любого типа.

Тексты можно отсылать в форматах *.doc, *.rtf (Microsoft Word) или *.pdf (тогда обязательно с расшифровкой, сопровождающей pdf). При использовании авторских шрифтов большая просьба **пересылать тексты вместе со шрифтами.**

3. Любые комментарии к текстам приветствуются (см. также Приложения 1 и 2).

4. Непосредственно в текст можно включать разъяснения диалектных слов, устойчивых сочетаний и фразеологии (в круглых скобках сразу после слова или ФЕ).

5. Текстам можно давать названия (*5. О плетении корзин*).

6. Можно сопровождать РАСШИФРОВАННЫЕ тексты аудио- и видеоматериалами (обещают, что это сопровождение будет технически возможно выставить в корпус в ближайшем будущем). Также можно иллюстрировать тексты фотографиями, схемами и картами.

Для передачи аудио- и видеоматериалов вместо e-mail может быть предоставлен канал FTP.

7. Материалы могут быть как публиковавшимися ранее, так и неопубликованными.

При размещении на сайте НКРЯ опубликованных текстов **обязательно будет сохраняться ссылка на публикацию**. Если публикация доступна в интернете - можно дать гиперссылку.

8. Каждый текст должен сопровождаться «адресом» (см. Приложение 1).

9. Лица, предоставившие тексты, могут принять участие в метаразметке этих текстов (см. Приложение 2)

10. Если у вас уже есть собственный сайт, содержащий тексты и / или описание говоров, в НКРЯ может быть помещена ссылка на этот сайт, что, без сомнения, во много раз увеличит его посещаемость.

Например: Тамбовские говоры. Фонохрестоматия:
http://93.186.97.70:81/kraeved/upload/dir/bibl/govor

11. Предполагается давать отсылки к сайтам, которые могли бы заинтересовать диалектологов:

– сайт «Информационный центр "Русская Диалектология"» (Институт Русского языка им. В.В. Виноградова РАН) (созданный «для обмена информацией между различными коллективами российских и зарубежных диалектологов, обеспечения планомерной и целенаправленной исследовательской и полевой работы в области диалектологии»): http://www.ruslang.ru/agens.php?id=rus_dialectology

– Сайт «Фонетика русских диалектов» (МГУ имени М.В. Ломоносова): http://dialect.philol.msu.ru/index.php

– Сайт «Школьный диалектологический атлас "Язык русской деревни"»: http://gramota.ru/book/village

– Сайты, на которых размещены общедоступные электронные версии диалектных словарей (отдельных выпусков словарей). Например, «Словарь русских народных говоров» (СРНГ): http://iling.spb.ru/vocabula/srng/srng.html)

12. Все рекомендации, советы и пожелания по обустройству сайта «Корпус Диалектных Текстов» в НКРЯ просим присылать по адресу: kacza@yandex.ru Качинской Ирине Борисовне

Приложение 1.

Адрес-сопровождение к тексту (заполнение обязательно)

1. Лицо, предоставившее текст:

ФИО _____

информация для связи: адрес, телефон, e-mail _____

(+ научное звание, должность, место работы) _____

2. Кем производилась запись _____

3. Место записи (область, район, населенный пункт)_____

4. Время записи (год)_____

5. Сведения об информанте_____

1) ФИО_____

***** в некоторых случаях по этическим соображениям при публикации на сайте фамилия информанта может быть сокращена до одной буквы – **на необходимость этого должно указать лицо, предоставившее текст**.

2) год рождения (возраст)_____

3) место рождения (местный, приезжий, откуда приехал, как давно живет в данном нас. пункте)_____

4) образование_____

5) профессия, род занятий (обычно их несколько)_____

6) Другие сведения _____

муж/жен – если это невозможно установить по имени-фамилии;

конфессия – если это важно (напр., для старообрядцев) и проч.

6. Где публиковался текст (если публиковался)_____

Выходные данные:

Автор публикации. Название статьи (книги). Место издания. Год издания. Номера страниц.

(*Петрова А.К.* Образцы говора юго-восточной зоны архангельского диалекта (Вилегодский р-н, д. Павловск) // Русские диалекты: история и современность. Вопросы русского языкознания. Вып. VII. М., 1997. С. 289-310)

7. Место хранения записи (научное учреждение, факультет, кафедра, фонд, личный архив)_____.

Приложение 2.

Сведения для метаразметки (заполнение желательно)

1. Тематика

Хорошо бы учитывать вопросник Лексического атласа русских народных говоров (ЛАРНГ), можно приводить номера вопросов, особенно если текст записывался при сборе сведений для ЛАРНГ.

2. Фонетические особенности говора

(тем более необходимо указывать, если текст оказался записан «в орфографии» или приближен к ней; если территория не попала в зону обследования ДАРЯ).

I. Вокализм

I.1. Ударный вокализм. Особенности ударного вокализма

поведение фонем /а/, /е/, /е-закрытого (старого ятя)/

а) перед твердым согласным,

б) между мягкими согласными.

I.2. Безударный вокализм 1-го предударного слога после твердого согласного:

Оканье, тип оканья (полное – неполное);

аканье, тип аканья (недиссимилятивное – диссимилятивное).

I.3. Безударный вокализм 1-го предударного слога после мягкого согласного.

I.4. Безударный вокализм непервого предударного после твердых / после мягких согласных (*заударное ёканье*).

+ другие особенности

II. Консонантизм

Особенности в произношении согласных:

– /г/, /в/, /л/, /аффрикат ц, ч/, /долгих шипящих – глухой и звонкой/, /j/,

– регулярное / нерегулярное изменение / упрощение групп согласных

+ другие особенности

3. Особенности грамматики

4. Синтаксические особенности

Особые синтаксические конструкции

– именительный прямого дополнения: *Косить трава*;

– безличные конструкции с причастиями: *У волков нахожено*; *Волками нахожено;*

– перфектные конструкции с деепричастием в качестве предиката *Он убежамши*;

– плюсквамперфект *Была болела, теперь поправилась*;

– особое предложное управление: *до председателя, о дорогу, возле магазин.*

+ другие синтаксические особенности

5. Другие особенности говора.

***

Благодарим всех диалектологов и фольклористов, уже предоставивших материалы для Корпуса – если вы еще не увидели на сайте свои тексты, они там непременно появятся. На всякий случай их можно продублировать по адресу: kacza@yandex.ru

Обращаемся также ко всем заинтересованным лицам: материалы, десятилетиями хранящиеся на кафедрах институтов, университетов и в частных собраниях, материалы, которые так давно ждут часа своей публикации, могут наконец быть опубликованы, дойти до читателя и исследователя. Мы все заинтересованы в пополнении Диалектного подкорпуса НКРЯ – пополнении как можно более быстром, значительном по объему и географической представленности. ЧАС НАСТАЛ!