ДИАЛЕКТНЫЙ ПОДКОРПУС СЕГОДНЯ[1]

И. Б. Качинская (kacza@yandex.ru**) Московский государственный университет им. М. В. Ломоносова**
Д. В. Сичинава (mitrius@gmail.com**) Институт русского языка им. В. В. Виноградова РАН**

1. Концепция Корпуса диалектных текстов

Корпус диалектных текстов на сайте НКРЯ существует с 2005 года http://www.ruscorpora.ru/search-dialect.html. Уже с самого начала отмечались особенности диалектной грамматики, приводился «паспорт» записи (место, время, данные об информанте и лице, предоставившем текст), имелись указания на жанрово-тематическую отнесенность текстов, на «диалектность» значений отдельных слов [Летучий 2005; 2009]. Вместе с тем тексты подавались в орфографизированном виде, не включалось толкование значений лексем.

За истекший период концепция подкорпуса существенно изменилась [Качинская 2009; 2011; Сичинава, Качинская 2014]. Возникла идея нового стандарта подачи как самих текстов, так и степени лингвистической обработанности каждой лексемы.

1.1. В настоящее время пользователю предоставлена возможность работать не только с фрагментами текстов (как во многих других подкорпусах, материалами для которых послужили тексты художественной литературы или СМИ), но и, если необходимо, с цельными текстами.

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). Туда входят полевые записи, аудиорасшифровки, тексты из малотиражных сборников и изданий, войдут также видеорасшифровки и тексты из хрестоматий. Мы надеемся, что со временем этот корпус станет репрезентативным собранием диалектных текстов и будет одним из самых посещаемых и востребованных пользователями. По материалам подкорпуса уже ведутся научные исследования [Пурицкая 2012].

1.2. Текст предоставляется пользователю в том виде, в котором он был первоначально записан, в том числе в фонетической транскрипции с сохранением ударений. Это особенно важно, т.к. диалектные тексты, особенно в транскрипционной записи, выпускаются в основном региональными

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

университетами небольшими тиражами и малодоступны даже в своей «книжной» ипостаси.

1.3. Все тексты 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).

1.4. Поиск может осуществляться по самым разным критериям.

На уровне **метаразметки** определяются:

1) география записи (с учетом области и района)
2) тематика
3) жанр
4) год записи
5) период истории, время описываемых событий
6) некоторые фонетические особенности
7) сведения об информантах (пол, возраст, образование, конфессия и нек. др.).

На уровне отдельной лексемы:

8) грамматика (с учетом диалектных особенностей)
9) семантика (не везде).
10) возможен поиск лексемы, включая ее словообразовательные варианты и синонимы (не везде)

2. Среда «Рабочее место диалектолога»

Для ответов на запросы по грамматике, лексике, семантике, тематике и проч. текст должен быть первоначально размечен. Для этого создан особый системный продукт: среда «Рабочее место диалектолога» (РМ, автор Т. А. Архангельский). В РМ осуществляется разметка как на уровне слова, так и на уровне всего текста (метаразметка).

Метаразметка предлагается на трех уровнях:

1) Адрес-сопровождение.
2) Фонетическая.
3) Диалектная текстовая.

Каждый из уровней представляет собой развертку в таблицу.

2.1. **Адрес-сопровождение**

В этой таблице отмечаются:

- Название текста
- Лицо, предоставившее текст (ФИО, информация для связи, научное звание, должность, место работы)
- кем производилась запись
- место записи (область, район, нас. пункт)
- год записи
- сведения об информанте (ФИО, возраст, год рождения, место рождения, образование, профессия / род занятий)
- где публиковался текст (если публиковался)
- место хранения записи

2.2. **Фонетическая метаразметка**

2.2.1. Пока что специально отмечаются лишь немногие фонетические диалектные особенности:

- в области ударного вокализма: позиционные чередования гласных после мягких согласных на уровне «старого ятя» и <а>;
- в области безударного вокализма: оканье и аканье (включая указания на неполное оканье или диссимилятивное аканье);
- в области консонантизма: отмечаются <г> взрывной или фрикативный и цоканье.

2.2.2. Пользователю представляется текст в фонетической транскрипции в двух вариантах: Текст-1 (Т-1) соответствует первоначальной записи, т.е. подается так, как он был представлен держателями текстов (в фонетической транскрипции или в орфографизированном виде); Текст-2 (Т-2) - в «облегченной», унифицированной транскрипции. Т-1 и Т-2 подаются в шрифтах Unicode. Если первоначальный текст существовал в других шрифтах, он переводится в Unicode.

Т-2 создается автоматически из Т-1 с помощью специально подготовленной программы **Детранскриптор-1** и практически не требует ручной доработки.

Соответствия Т-1 и Т-2:

Т-1	Т-2
в'и́т'ер, н'а́н'к'а, н'осу́	*ви́тер, ня́нькя, нёсу́*
кон', з'ем'е́л'к'е	*конь, земе́льке*
jéс'т', najдý, см'ujóца, ружйó	*йе́сьть, пайдý, смиёца, ружьё*

Некоторая сложность возникает при использовании в Т-1 знака *Ь*, обозначающего редуцированный гласный переднего ряда. Выход из положения видится следующий: в Т-1 текст отается с использованием знака *Ь* в качестве гласного. В Т-2 гласный [ь] заменяется на знак [и], а знак *Ь* используется как «мягкий знак» (так же, как в принятой орфографии).

Т-1: *в'ьч'ьре́лъ*; Т-2: *вичире́лъ*

Т-1 и Т-2 выступают как параллельные тексты.

2.2.3. 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.

Таким образом, если нет никаких отметок, то запись подана в транскрипции. Если стоит отметка *Орфографизированная запись,* значит, запись подана либо в орфографии, либо в орфографизированном виде - т.е. могут отмечаться какие-то отдельные диалектные особенности в области произношения на фоне общей установки на орфографию. Например, может быть отмечена утрата затвора у <ч>, так наз. «шоканье» (*девощка* = девочка = [дéвъщка]) при отсутствии отображения аканья в акающем говоре; могут быть отмечены какие-то особенности грамматики (*они стоя́ть* = они стоят = [ани стая́ть]).

Если запись первоначально была орфографической, то Т-1 и Т-2 совпадут.

Подача текстов в орфографизированном виде связана с достаточно устойчивой традицией публикации диалектных текстов. Иногда это делалось из-за недостатков полиграфической базы, сложности печати, но порой - для лучшего

понимания таких текстов читателями. Вместе с тем даже фольклористы, публикуя тексты в «орфографии», часто не могут устоять от соблазна показать особенно яркие черты фонетики (*Ванькя клюшник злой разлушник...*)

2.3. **Диалектная текстовая метаразметка** включает в себя 3 подуровня:

- жанр (тип) текста;
- место и время описываемых событий;
- тематика текста.

Каждый подуровень имеет свое деление.

2.3.1. **Жанр (тип) текста** делится на 3 категории: 1.1. Устные нефольклорные тексты; 1.2. Письменные нефольклорные текты; 1.3. Фольклорные тексты. При выборе типа текста возникает возможность выбора конкретного жанра. Например, в **1.1. Устные нефольклорные тексты** включены следующие жанры: диалог; рассказ; пересказ (сна / разговора / фильма / телепередачи); ответы по вопроснику; спор. **1.2. Письменные нефольклорные тексты** включают жанры: автобиография; воспоминания / мемуары; личное письмо; поздравление; личный дневник / записная книжка; служебная / докладная записка; просьба, прошение; справка; заявление; характеристика; художественный текст, созданный носителем диалекта (в том числе поэтический).

Пока предпочтение при включении 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 оказываются пословицы и поговорки, загадки, частушки и колыбельные песни и проч.). В то же время в банке диалектных текстов уже есть как письменные фольклорные тексты («песенники» или заговоры, записанные самими носителями), так и письменные нефольклорные (письма, мемуары, дневники, записные книжки, стихи и проч.).

2.3.2. **Место и время описываемых событий** во многом повторяет подобную таблицу в других корпусах НКРЯ и включает в себя следующие разделы:

- Доисторический период
- За пределами России
- Ирреальный (фантастический) мир
- Россия

Выделяются следующие исторические периоды, значимые как для русской деревни в целом, так и для наших информантов, их личной биографии (с XX века):

- до XVIII века
- XIX век
- 1900-1914
- 1914-1920
- 1920-1930-е (коллективизация)
- от коллективизации до войны 1941-1945
- война 1941-1945
- послевоенное восстановление 1945-1953
- 1950-1980-е
- постсоветский период

2.3.3. **Тематика текста** включает достаточно много разделов, однако требует кардинальной переработки. Первоначально тематическое дерево было

ориентировано на вопросники Лексического атласа русских народных говоров (ЛАРНГ), разработанные в ИЛИ РАН. Но вскоре выяснилось, что дробная семантическая классификация, направленная на **слово**, не вполне годится для определения тематики **текста**.

3. Особенности грамматической разметки

3.1.1. Грамматическая разметка диалектных текстов также осуществляется в РМ. Так как тексты в Диалектном подкорпусе поданы либо в транскрипции, либо в орфографии с сохранением некоторых фонетических и грамматических особенностей, возникла проблема их грамматического распознавания. Было принято решение к каждому тексту делать орфографический «подстрочник», создавать так наз. Текст-3 - орфографизированный вариант диалектного текста, по которому сможет работать стандартный грамматический анализатор, применяемый во всех корпусах НКРЯ. Для этой цели был создан **детранскриптор-2**, переводящий Т-2 (или Т-1) в Текст-3 (Т-3). Орфографический «подстрочник» требуется доводить до уровня орфографии вручную. Детранскриптор-2 основан на серии последовательных замен некоторых буквенных сочетаний, отличающих русскую фонетику от орфографии, отдельной строкой в него могут быть вписаны любые условия («*пробел – ана – пробел*» заменяется на *она*). Детранскриптор можно постоянно унифицировать, вписывая условия для наиболее часто повторяющихся замен. Унификация детранскриптора-2 увеличит степень «орфографизированности» текста, но не отменит ручной доводки. В РМ в Т-3 специальной программой осуществляется проверка стандартной орфографии, цветной меткой помечаются слова, с орфографией не совпадающие. Это или ошибки разметчика, которые легко исправить, или диалектные слова (словоформы), отсутствующие в основном словаре.

Тексты 1, 2 и 3 выровнены по предложениям в формате XML, как в Параллельном корпусе НКРЯ.

Пример выровненного размеченного текста:

<para>

<se format="1">/ на / дифчёнкъ // а радить будить/ а рибятúшки будут с красными въласами и ех замуш нихтó ни вазьмёть // ну / тёть Шуръ Кьралёвъ ұъварит / ну куды ш паедим?</se>

<se format="2">/ <w><ana lex="на" gr="PART"></ana><ana lex="на" gr="PR"></ana>на</w> / <w><ana lex="девчонка" gr="S,anim,f,sg,nom,1decl"></ana>дифчёнкъ</w> // <w><ana lex="а" gr="INTJ"></ana><ana lex="а" gr="S,0"></ana><ana lex="а" gr="CONJ"></ana>а</w> <w><ana lex="родить" gr="V,pf,inf,act,2conj"></ana><ana lex="родить" gr="V,pf,tran,inf,act,2conj"></ana><ana lex="родить" gr="V,ipf,tran,inf,act,2conj"></ana>радить</w> <w><ana lex="будить" gr="V,ipf,tran,inf,act,2conj"></ana>будить</w>/ <w><ana lex="а" gr="INTJ"></ana><ana lex="а" gr="S,0"></ana><ana lex="а"

gr="CONJ"></ana>а</w> <w><ana lex="ребятишки" gr="S,anim,pl,nom,pltantum"></ana>рибятúшки</w> <w><ana lex="быть" gr="V,pf,intr,indic,act,fut,3p,pl,1conj"></ana>будут</w> <w><ana lex="с" gr="PR"></ana><ana lex="с" gr="S,0"></ana>с</w> <w><ana lex="красное" gr="S,inan,n,pl,ins"></ana><ana lex="красный" gr="S,anim,m,pl,ins,adjdecl"></ana><ana lex="красный" gr="A,pl,ins,plen"></ana>красными</w> <w><ana lex="волосы" gr="S,inan,pl,ins"></ana><ana lex="волос" gr="S,inan,m,pl,ins,2decl"></ana>въласами</w> <w><ana lex="и" gr="INTJ"></ana><ana lex="и" gr="PART"></ana><ana lex="и" gr="S,0"></ana><ana lex="и" gr="CONJ"></ana>и</w> <w><ana lex="их" gr="APRO"></ana><ana lex="они" gr="SPRO,anim,pl,gen"></ana><ana lex="они" gr="SPRO,anim,pl,acc"></ana>ех</w> <w><ana lex="замуж" gr="ADV"></ana>замуш</w> <w><ana lex="никто" gr="SPRO,anim,m,sg,nom"></ana>нихтó</w> <w><ana lex="ни" gr="PART"></ana><ana lex="ни" gr="CONJ"></ana>ни</w> <w><ana lex="взять" gr="V,pf,tran,indic,act,fut,3p,sg,1conj"></ana>вазьмёть</w> // <w><ana lex="ну" gr="INTJ"></ana><ana lex="ну" gr="PART"></ana>ну</w> / <w><ana lex="тетя" gr="S,anim,f,pl,gen,1decl"></ana><ana lex="тетя" gr="S,anim,f,pl,acc,1decl"></ana>тёть</w> <w><ana lex="шура" gr="S,persn,anim,pl,gen,1decl"></ana><ana lex="шура" gr="S,persn,anim,pl,acc,1decl"></ana>Шуръ</w> <w><ana lex="королев" gr="S,famn,anim,m,sg,gen,famdeclm"></ana><ana lex="королев" gr="S,famn,anim,m,sg,acc,famdeclm"></ana><ana lex="королева" gr="S,famn,anim,f,sg,nom,1decl"></ana><ana lex="королева" gr="S,anim,f,sg,nom,1decl"></ana><ana lex="королев" gr="S,inan,m,sg,gen,famdeclm"></ana>Къралёвъ</w> <w><ana lex="говорить" gr="V,ipf,tran,indic,act,praes,3p,sg,2conj"></ana>ӷъварит</w> / <w><ana lex="ну" gr="INTJ"></ana><ana lex="ну" gr="PART"></ana>ну</w> <w><ana lex="куда" gr="ADVPRO"></ana>куды</w> <w><ana lex="ш" gr="S,0"></ana>ш</w> <w><ana lex="поехать" gr="V,pf,intr,indic,act,fut,1p,pl,1conj"></ana><ana lex="поехать" gr="V,pf,intr,imper,act,1p,pl,1conj"></ana>паедим</w>?</se>

<se format="3">/ <w><ana lex="она" gr="SPRO,anim,f,sg,nom"></ana>на</w> / <w><ana lex="девчонка" gr="S,anim,f,sg,nom,1decl"></ana>девчонка</w> // <w><ana lex="а" gr="CONJ"></ana>а</w> <w><ana lex="родить" gr="V,pf,tran,indic,act,fut,3p,sg,2conj,dialflex"></ana>родит</w> <w><ana lex="быть" gr="V,pf,intr,indic,act,fut,3p,sg,1conj,dialflex"></ana>будет</w>/ <w><ana lex="а" gr="CONJ"></ana>а</w> <w><ana lex="ребятишки" gr="S,anim,pl,nom,pltantum"></ana>ребятишки</w> <w><ana lex="быть" gr="V,pf,intr,indic,act,fut,3p,pl,1conj"></ana>будут</w> <w><ana lex="с" gr="PR"></ana>с</w> <w><ana lex="красный" gr="A,pl,ins,plen" lex_ref="рыжий"></ana>красными</w> <w><ana lex="волосы" gr="S,inan,pl,ins"></ana>волосами</w> <w><ana lex="и" gr="CONJ"></ana>и</w> <w><ana lex="они" gr="SPRO,anim,pl,acc,dialstem,dialform"></ana>их</w> <w><ana lex="замуж" gr="ADV"></ana>замуж</w> <w><ana lex="никто" gr="SPRO,anim,m,sg,nom"></ana>никто</w> <w><ana lex="не" gr="PART"></ana>не</w> <w><ana lex="взять" gr="V,pf,tran,indic,act,fut,3p,sg,1conj,dialflex"></ana>возьмёт</w> // <w><ana

lex="ну" gr="INTJ"></ana><ana lex="ну" gr="PART"></ana>ну</w> / <w><ana lex="тетя" gr="S,anim,f,sg,voc,1decl"></ana>тёть</w> <w><ana lex="шура" gr="S,persn,anim,sg,nom,1decl"></ana>Шура</w> <w><ana lex="королева" gr="S,famn,anim,f,sg,nom"></ana>Королёва</w> <w><ana lex="говорить" gr="V,ipf,tran,indic,act,praes,3p,sg,2conj"></ana>говорит</w> / <w><ana lex="ну" gr="PART"></ana>ну</w> <w><ana lex="куды" gr="ADVPRO" lex_ref="куда"></ana>куда</w> <w><ana lex="ж" gr="PART"></ana>ж</w> <w><ana lex="поехать" gr="V,pf,intr,indic,act,fut,1p,pl,1conj"></ana>поедем</w>?</se>
</para>

3.1.2. После создания орфографического «подстрочника» осуществляется **грамматическая разметка** Текста-3 стандартным грамматическим анализатором. Дальше текст обрабатывается вручную:

1) снимается грамматическая **омонимия** (как в Основном корпусе),

2) отмечаются **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 (глагол, существительное, прилагательное, местоимение, числительное).

3.2. Так как грамматические особенности русских говоров достаточно хорошо изучены, возникла идея использовать таблицы с реальными диалектными аффиксами, формами и проч. Однако при работе с конкретными текстами каждый раз выяснялось, что списки введенных аффиксов оказывались неполными. Чтобы не множить сущности, было принято решение вернуться к обобщенным указаниям на диалектные черты отдельных грамматических категорий: рода, числа, падежа, склонения, вида, переходности, возвратности и проч.

Для каждой изменяемой части речи предусмотрена возможность указывать диалектные особенности на многих уровнях. Приведем некоторые примеры.

3.3. Диалектные особенности у **существительных** могут быть связаны с категориями рода, числа, типа склонения, одушевленности; они могут иметь иные нежели в ЛЯ, падежные окончания.

3.3.1. Категория **рода**: *берлог* (= берлога)*, пуза* (= пузо)*, литра* (= литр), *девчонко* (= девчонка)*, яблок* (= яблоко), *зайко, дедушко* и т.п. При указании на диалектные особенности рода принимаются в расчет два критерия: морфологический - по типу склонения, синтаксический - по согласованию с адъективами или с глаголами в прош. вр. (*плачь у ней была такая*), в том числе в случаях рассогласования, т.е. когда синтаксический и морфологический критерии не совпадают (*така хороша девчёшко*, *парнáя мълако*). Если особенности рода сопровождаются особенностями в типе склонения, то отмечается одновременно и «диалектный род», и «диалектное склонение» (*такую колечку*).

3.3.2. Изменение типа **склонения** (или особенности склонения) могут быть связаны как с распадением категории среднего рода - при ориентации сущ. среднего рода на женский (*мою колечку, в девятую царству*) или мужской (*вкусный яблок*), так и с ориентацией 3 скл. на 1-е: *на пече, ночей, осеньей.* В этом

случае как диалектные особенности отмечаются «диалектное склонение» и «флексия» (dialflex).

3.3.3. **Диалектная флексия (dialflex)** у имен существительных может указывать на разные вещи. Так, у существительных 1 скл. в части говоров могут совпадать окончания Д=П и Р: *к козы́, в реки́, в Москвы́, на войны́.* Есть говоры, в которых Р совпадает с Д=П (окончание <*е*> из «ятя»): *у жене́, без сестре́,* в безударной позиции в акающих говорах: *у баби* (<*у бабе*>), *блис школи* (<*близ школе*>). У существительных 2 скл. особенности окончаний могут встречаться в Р. (говоры с расширенной сферой окончания -*у*: *без чаю, без мужику, из городу*) или в П. (говоры, в которых окончания П. -*е* и -*у* распределены иначе, чем в ЛЯ: *в лесе, в глазе, в шкафе, на береге* или *на коню*; в П. встречается грамматикализованное ок. -*и*: *на кони*). Особые окончания могут встретиться у так наз. разносклоняемых существительных (существительных с нерегулярным склонением): слова на -*мя, мать, дочь, свекровь, день, путь* и некоторых других, а также во множественном числе (*без бабов, месяцей, лопаткима*). Если исследователя интересует диалектные грамматические особенности определенной лексемы или определенного падежа, достаточно задать в поиске параметры падежа, числа, типа склонения + диалектные особенности.

3.4. Диалектные особенности **глагола** могут быть связаны с категориями вида, переходности, возвратности, наклонения, типов спряжения; глаголы в русских народных говорах могут иметь специфические окончания в личных формах, особые основы, иные формы инфинитива, императива, причастий, деепричастий, употребляться в конструкциях, связанных с иным временем (перфект, плюсквамперфект). Во всплывающей таблице «диалектные особенности» для глагола отмечаются как «диалектное»: **основа, суффикс, флексия, форма, вид, переходность, возвратность, время**.

3.4.1. Иная, чем в ЛЯ, **основа** глагола **(dialsteam)** может быть связана с принадлежностью глагола к регулярному или нерегулярному классу, т.е. с иным соотношением основ настоящего и прошедшего времени (инфинитива): наст. время может ориентироваться на прошедшее (*торговать ~ торгует, ездил ~ ездию, мыл ~ мыет*); прошедшее – ориентироваться на настоящее (*жму ~ жмал, трёт ~ тратъ, умрёт ~ умрал*). Это явление может считаться «архаическим» или «новаторским» (*гоню ~ гонить / гнать, мяукать / мяучить ~ мяукает / мяучит, чихать ~ чишу / чихаю*, *полоскать ~ полощу / полоскаю*). Диал. основа может быть связана с процессом выравнивания (унификации) основ личных форм глагола в настоящем - простом буд. времени: *просю, любю, бегёшь, бежат, пекёт, пекот* (= печет).

3.4.2. Диалектный **суффикс (dialsfx)** отмечается при особом **императив**е: *поста́ви* (= поставь), *удари* (= ударь), *посоль* (= посоли) – т.е. учитывается и нулевой суффикс. Если форма императива возникает от иной, нежели в ЛЯ, основы (*доедь, ехай, пахай*), то отмечается не dialsfx, а dialform (диалектная форма). Таким образом, исследователю для того, чтобы получить сведения об особом императиве, достаточно задать в поиске императив + диалектные особенности.

3.4.3. Именно суффиксом отличаются в разных говорах формы **инфинитива**: «архаический» -*ти* безударный (*гуляти, смотрети*); «новаторский» -*ть* с утратой гласного (*потрясть* = потрясти, *подместь* = подмести); «новаторский» -*чи* в глаголах с основой на заднеязычный (*пекчи* = печь, *стерегчи* = стеречь). В этих

случаях отмечается диалектный суффикс (dialsfx). В случаях, когда встречается унификация суффикса инфинитива (*идти́ть, прийти́ть, пойди́ть*), отмечается не dialsfx, а dialform (диалектная форма). Если особая форма инфинитива связана с переходом / непереходом глагола в регулярный класс, т.е. с изменением соотношения основ непрошедшего / прошедшего времени: *жмать* = жать (руку), *жнить* = жать (траву), *бости* = бодать, *ноять* = ныть, *бечь* = бежать, то одновременно отмечается dialform (диалектная форма) и dialsteam (диалектная основа).

3.4.4. Из четырех видов **причастий** в говорах, как правило, встречаются лишь страдательные причастия прошедшего времени (чаще в краткой форме), которые образуются с помощью тех же суффиксов, что и в ЛЯ (*-енн-, -нн-, -т-*), однако в разных лексемах эти суффиксы могут быть распределены иначе, чем в ЛЯ (*убратый* = убранный, *датый* = данный, *расколонный* = расколотый). В этих случаях отмечается диалектный суффикс (dialsfx). Если причастия образованы от иных основ (*даденный*), в таком случае отмечается dialsteam (диалектная основа) и/или dialform (диалектная форма).

3.4.5. Как правило, в говорах встречается иной набор суффиксов **деепричастий**: *-вши, -лши, -мши, -тши, -чи* (*выпивши, выпимши, выпилши, пришёдци*). В этом случае отмечаются диалектный суффикс (dialsfx) и диалектная форма (dialform). Часто такие деепричастия используются в функции предиката и имеют значение перфекта (*Он пришоццы. Корыто было пропалшы*) – тогда кроме диалектного суффикса и диалектной форма должно быть отмечено диалектное время (dialtense). Таким образом, если пользователь хочет найти деепричастия в функции перфекта, он задает запрос на деепричастия + диалектное время.

3.4.6. Если для установления непереходности диалектного глагола требуется большой иллюстративный материал, то **переходность** определяется очень хорошо (*сходила его*). В таком случае отмечается диалектная переходность (dialtrans).

3.4.7. В графу **диалектная возвратность (dialrefl)** попали самые разные случаи, характеризующие особенности возвратных глаголов. Так, в говорах на месте возвратного глагола (в ЛЯ) может оказаться невозвратный (*смеять*), а на месте возвратного – невозвратный (*дрожаться, дежуриться*). Возможно присоединение постфикса *-ся* не только к основе на согласный, но и на гласный (архаические формы: *умывалася, посмотрелися*). В постфиксе может присутствовать как мягкий *с*, так и твердый (*умывалас, умывалса*). Гласный постфикса может быть не только *а*/*я*, но и *и*/*ы, е*/*э, ё*/*о* (*умывался, умывалсе, умывалси, умывалсё; умывалса, умывалсэ, умывалсы, умывалсо*). Если пользователь хочет найти невозвратные глаголы на месте возвратных, он задает запрос по диалектной возвратности, при этом отмечает отсутствие *-ся* в конце слова. Ассимилятивные процессы на стыке основы и постфикса (*смеёсся, смеёшиша*) являются фонетическими и в особенностях грамматики не учитываются.

3.4.8. **Диалектные окончания (dialflex)** характеризуют глаголы 3 л. ед. и мн. ч., сюда также попадают самые разные случаи: ударные окончания без перехода *е* > *о* (*идо́шь*), конечное *-ть* в 3 л. ед. и мн. ч. (*растёть, растуть*), формы без *-т* (*идё*), общее спряж. (*смотрют*), так наз. 3-е спряж. (*играт*) и проч. Думается, что специалисту в этом достаточно легко разобраться, особенно при возможности учитывать географические фильтры. Например, если исследователь хочет найти особенности спряжения и ищет «общее спряжение» (*любют*), он задает в поиске

«наст.вр. 3 л. мн. ч. + dialflex+ ок. *-ут, -ют*». Если его интересует 3-е спряжение, характерное для северных и среднерусских говоров (*он гулят*), лучше включать географические фильтры и указывать 2 и 3 л. ед. и 1, 2 и 3 л. мн. ч. наст.вр. + диал.флексия.

3.4.9. 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. В.-Р. ед., или *ихной, ихний, ихой* (= их), *тэй* / *тый парень* (= тот), *оне, оны* (= они), сравн. ст. прилаг. / наречий - *первеющий* и проч.

4. Развитие корпусной диалектологии на текущем этапе

4.1. Создание диалектных корпусов – дело во многом новое, особенно общедоступных корпусов. Назовем известные нам продукты, доступные через интернет, в которых отражены сведения о русских говорах или представлены диалектные тексты:

- сайт «Школьный диалектологический атлас "Язык русской деревни"» (включающий карты и подробные комментарии к ним): http://gramota.ru/book/village (ИРЯ РАН, Москва).
- Фонетика русских диалектов: http://dialect.philol.msu.ru/index.php (МГУ имени М.В. Ломоносова, Москва).
- «Электронная библиотека русских народных говоров» (собраны материалы экспедиций в различных говорах Европейской части России): http://dialekt.rx5.ru/index.html (Казанский [Приволжский] федеральный ун-т, Казань).
- «Лингвогеографическая система «Диалект»»: http://lgw2.udsu.ru:9001/ (Удмуртский ун-т, Ижевск).
- «Региональная этнолингвистика» (материалы по Кубанским говорам): http://www.ethnolex.ru/ (Кубанский государственный университет, Славянск-на-Кубани).
- Тексты из Шатурского р-на Московской обл. и Харовского р-на Вологодской обл. в рамках проекта «Электронные базы данных по русским народным говорам»: http://starling.rinet.ru/cgi-bin/main.cgi?root=ruscorpora&encoding=utf-rus (ИРЯ РАН, Москва, авторы - С. А. Крылов и А. В. Тер-Аванесова)
- тексты и база по говору д. Пушкино Устьянского р-на Архангельской обл. (Ustja River Basin Corpus Query interface): http://www.slavist.de/Pushkino/login.php (Национальный исследовательский университет «Высшая школа экономики», Москва, и институт славистики Бернского университета, Берн, Швейцария; Научно-учебная группа «Корпусное исследование вариативности в речи носителей говора Устьянского района Архангельской области», рук. М. А. Даниэль и Р. фон Вальденфельс).

4.2. Во многих научных центрах (как правило, организованных на базе университетов) ведется активная работа по созданию и совершенствованию

диалектных корпусов (пока еще без выхода в Интернет): «Саратовский диалектологический корпус» (http://sarteorlingv.narod.ru/dialekt/kru4kova-goldin.html) создан по материалам трех русских говоров - двух южных и одного северного (Центр изучения народно-речевой культуры Саратовского государственного университета им. Н. Г. Чернышевского, руководители: проф. В. Е. Гольдин и проф. О. Ю. Крючкова; http://www.sgu.ru/structure/philological/narrech). Материал более чем из ста говоров Архангельской области содержится в Корпусе «Электронная картотека "Архангельского областного словаря"» (МГУ имени М. В. Ломоносова, Москва) - 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 [Русская устная речь 2011; Юрина 2011].

4.3. Сбор диалектных текстов для выставления их на сайте НКРЯ не должен и не может 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 диалектных текстов НКРЯ. С другой стороны, нельзя игнорировать уже имеющиеся образцы записи народных говоров, которые оказались далеки от нашего «стандарта». Главное отличие Диалектного подкорпуса 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, будут основываться на жанрово-тематической разметке.

В Корпусе диалектных текстов НКРЯ предполагается давать ссылки на существующие корпуса, имеющие сайты в интернете.

4.4. Если грамматическая разметка дается в Диалектном подкорпусе для каждого слова, то семантика слова подается лишь в том случае, если тексты сопровождаются словарем или дается ссылка на имеющиеся словари. В некоторых случаях семантика уточняется за счет указанных синонимов (в поле lexref). Конечно, этого недостаточно даже для лингвиста-недиалектолога, не говоря уже о пользователях-нелингвистах, не знакомых с говором. Выход видится в том, чтобы в Диалектном подкорпусе ввести ссылки на диалектные словари, например на «Словарь русских народных говоров» http://iling.spb.ru/vocabula/srng/srng.html (ИЛИ РАН, Санкт-Петербург); «Архангельский областной словарь»: http://www.philol.msu.ru/~dialectology/dictionary/ (МГУ имени М. В. Ломоносова, Москва). К сожалению, подавляющее количество диалектных словарей не имеет в интернете официальных адресов (сайтов).

4.5. В Диалектном подкорпусе возможно включать любого рода комментарии к текстам: фонетические, грамматические, семантические, экстралингвистические (предоставить этнографическую, этнокультурную информацию), возможно сопровождение текстов фотоматериалами. 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.

4.6. Свободное предоставление в интернете текстов, записанных в русских народных говорах, их грамматическая, семантическая и метатекстовая характеристика позволит специалистам-диалектологам, другим лингвистам и нелингвистам, филологам, историкам, культурологам, этнографам и всем, кто интересуется народным русским словом, обращаться к подкорпусу 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.

Литература

Качинская И.Б. 2009. Корпус Диалектных Текстов в Национальном корпусе русского языка: состояние и перспективы // Лексический атлас русских народных говоров (Материалы и исследования). 2009. СПб. С. 57-68. (http://www.philol.msu.ru/~ruslang/pdfs/kachinskaya.i.b/19.pdf от 01.03.2014)

Качинская И.Б. 2011. Диалектный Подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место // В сб.: 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». Саратов, СГУ, 15-17 ноября 2010 г., ред. О.Ю. Крючкова, А.И.Буранова, В.Е.Гольдин, Л.В. Балашова. С. 245-255.

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

Летучий А. Б. 2009. Диалектный корпус: состав и особенности разметки // Национальный корпус русского языка: 2006-2008. Новые результаты и перспективы. СПб.: Нестор-История. С. 114-128.

Сичинава Д.В., Качинская И.Б. 2014. Корпус диалектных тестов в национальном корпусе русского языка: сегодняшнее состояние и перспективы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4-8 июня 2014 г.). Вып. 13 (20). – М.: Изд-во РГГУ, 2014. С. 593-600.

Пурицкая Е.В. 2012. Диалектный подкорпус национального корпуса русского языка как источник изучения лексической динамики диалекта // Северно-русские говоры. Вып. 12: межвуз. сб. / отв. ред. А.С. Герд. – СПб.: Изд-во С.-Петерб. ун-та, 2012. С. 14-22.

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». Саратов, СГУ, 15-17 ноября 2010 г., ред. О. Ю. Крючкова, А. И.Буранова, В.Е.Гольдин, Л.В. Балашова.

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 01.09.2014).