А. Б. Летучий

КОРПУС ДИАЛЕКТНЫХ ТЕКСТОВ: ЗАДАЧИ И ПРОБЛЕМЫ

Задачи диалектного корпуса

Диалекты русского языка давно стали предметом лингвистических исследований. В настоящее время проводятся регулярные экспедиции — в частности, нужно отметить экспедиции Института русского языка, Саратовского государственного университета. Кроме того, диалекты изучаются с точки зрения фольклористики и других областей. Тексты на диалектах публикуются с комментариями и описаниями разной степени подробности. Однако корпуса диалектных текстов с морфологической и метаразметкой до сих пор не существовало. В ближайшее время на сайте Национального корпуса русского языка будет размещён отдельный корпус диалектных текстов.

Такой корпус необходим сразу по нескольким причинам: во-первых, он позволит легко сравнивать диалекты с литературном русским языком: в частности, как мы увидим, в диалектном корпусе специально отмечаются отличия от литературного языка. Во-вторых, лексико-грамматический поиск, используемый в Национальном корпусе, позволит собирать материал для изучения грамматических свойств диалектов. Наконец, задавая подкорпус по своему желанию, можно сравнивать разные диалекты или разные тексты между собой.

Унификация текстов

Поскольку сбор и запись текстов осуществлялись в разное время и разными экспедициями, принятые системы записи и транскрипции достаточно сильно отличаются друг от друга. В частности, в первом собрании (юго-западные говоры) опускаются реплики интервьюера, которые включаются в текст в остальных собраниях. С другой стороны, во втором и третьем собрании последовательно маркируются паузы — как между предложениями, так и внутри них, — которые в первом корпусе никак не отмечаются. Тем не ме-

нее, для помещения текстов в корпус эти расхождения не слишком существенны.

Более существенно то, что в разных собраниях приняты разные правила обозначения редукции гласных. В первом используются только знаки, совпадающие с русскими буквами. Соответственно, из позиционных чередований гласных на письме обозначаются только те, которые существенно изменяют качество гласного (например, отмечается переход [е] в [и], но не [а] в [ъ]). Хотя эта транскрипция не слишком точна, в то же время она наиболее удобна для Корпуса: поскольку основные задачи создания диалектного корпуса — изучение проблем морфологии и словоупотребления, желательно отмечать только наиболее существенные фонетические изменения.

Напротив, во втором и третьем собрании текстов используется более точная фонетическая транскрипция: отмечаются изменения в заударных и предударных слогах и в безударных словах (например, *дъ**к*** ‘так’, *ъ**т**елиця* ‘отелится’) и др. Также отмечаются случаи изменения [и] в [ы] (ср. союз *и*). Не всегда ясно, отражают ли эти различия в записи реальные расхождения в вокализме, или некоторые говоры просто записывались более точно.

Также во втором и в третьем собрании используются расширенные наборы знаков для согласных, в частности, имеются специальные обозначения для двух видов «л» ([л] и [l]) и для билабиального звука [w]. В диалектном корпусе эти различия учитываться не будут, слова, содержащие [л] в литературном языке, будут записываться с ним и в диалектных текстах — главной целью создания корпуса диалектных текстов всё же не является фиксация фонетических особенностей диалектов. Кроме того, судя по всему, эти особенности фиксировались диалектологическими экспедициями с разной подробностью и разными способами, что мешает сравнению свойств фонетической системы диалектов. Нужно отметить, что, судя по всему, их распределение не вполне ясно (ср. *ц**é**w**o*** ‘целое’ и *бы**l**o* ‘было’).

Морфологически ориентированная стратегия

При разметке диалектных текстов, как и для литературных текстов, принимается морфологически ориентированная стратегия: отмечаются только те отличия от литературного языка, которые имеют отношение к грамматике или отражаются на грамматиче-

ских особенностях, в частности, (1) меняют тип словоизменения, (2) приводят к возникновению новых типов омонимии грамматических форм. Так, не учитывается редукция гласных в корнях: во-первых, те же типы редукции во многом происходят и в литературных текстах, во-вторых, она не меняет внешнего вида грамматических показателей.

С другой стороны, случаи типа окончания *-ой* в форме именительного падежа единственного числа прилагательных (*которой*, *хорошой*) в текстах отражаются. По всей вероятности, и эти изменения мотивированы фонетически: в северных диалектах имеет место «полное оканье», в результате чего окончания *-ый/-ий* могут переходить в *-ой*. Однако переход в окончании влияет на тип образования номинатива (он образуется по типу *плохой*) и на омонимию (возникает омонимия типа *жалосливой* — [им. п., ед. ч., м. р.] vs [твор. п., ед. ч., ж. р.]).

Изменения постфикса *-ся*, имеющего варианты *-се*, *-ся* и *-с'*, первый из которых в литературном языке отсутствует, вряд ли стоит учитывать при унификации: во-первых, формы на *-ся* в любом случае не омонимичны ни с какими другими, во-вторых, словоизменительный класс глагола не зависит от выбора варианта постфикса. Кроме того, по-видимому, изменения *-ся* регулируются всё же фонетическими и морфонологическими правилами, хотя и факультативными. Наконец, и в литературном языке суффикс имеет варианты, в некоторых случаях находящиеся в отношениях свободного варьирования.

Напротив, нулевая форма 3 л. настоящего времени требует отражения в морфологической разметке: в отличие от возвратного суффикса, здесь не вполне понятно, вызвано ли выпадение *-т* чисто фонетическими причинами, не слишком очевидно, являются ли две формы вариантами друг друга. Кроме того, случаи типа диалектного образования презенса обозначаются в корпусе пометой DIALMORPH («диалектная морфология»).

Таким образом, важны не источники и причины изменений, а их вес в грамматической системе и влияние на набор грамматических классов. В будущем, возможно, будут отражены наиболее частотные фонетические изменения, не влияющие на грамматические характеристики (например, оканье).

Существующие диалектные корпуса

Среди существующих корпусов текстов большинство использует материал литературных языков (ср. такие известные корпуса, как Британский национальный корпус, Тюбингенский корпус русских литературных текстов и др.). Тем не менее сегодня существует довольно много отдельных корпусов диалектных текстов, кроме того, некоторые языковые корпуса содержат диалектные подкорпуса.

В качестве примера можно привести Хельсинкский корпус (The Helsinki Dialect Corpus of British English).

Как и диалектный подкорпус Русского национального корпуса, данный корпус, как указано в работе [Peitsara, Vasko 2002], создан, прежде всего, для морфосинтаксических, а не для фонетических исследований (как известно, часто целью исследований диалектов становится именно изучение фонетической системы).

В корпусе принята следующая система подачи информации. Формы, совпадающие с литературными, записываются в виде текста, однако диалектные формы (в особенности допускающие неоднозначную интерпретацию) записываются в фонетической транскрипции — таким образом на них обращается внимание пользователя.

Пунктуация в диалектных текстах не обязательно следует литературным правилам — она служит прежде всего для того, чтобы дать некоторое представление о членении диалектной речи и облегчить читателю восприятие текстов.

В корпусе отмечаются паузы хезитации и фальстарты, дающие возможность проследить за выбором говорящим средств выражения.

С другой стороны, создаются корпуса диалектных текстов и для восточных языков, ср., например, корпус, созданный в Китае в рамках «Программы 863» (см. [Qian Yue Liang et al. 2000]). Этот корпус включает в себя большое количество аудиоматериалов, разделённых на шесть подкорпусов: «Корпус синтеза китайской речи», «Корпус материалов для просодического анализа», «Корпус смешанного чтения на китайском и английском языках», «Корпус распознавания китайской речи», «Корпус распознавания китайской речи по телефону» и «Диалектный корпус мандаринского китайского». Диалектный корпус включает данные, полученные от восьмисот волонтёров из четырёх городов, большинство из них среднего возраста. Каждый волонтёр в течение одного или двух

часов читал вслух заранее заготовленные материалы, в результате объём корпуса составил 40 Гб или 40 часов разговорной речи. Очевидно, что данный корпус ориентирован, прежде всего, на изучение фонетических особенностей китайских диалектов. Как кажется, недостатком является то, что корпус состоит из заранее созданных текстов — хотя их зачитывали носители диалектов, при зачитывании чужого текста особенности речи носителей могли нивелироваться из-за желания читать текст возможно более внятно, кроме того, ситуация чтения чужого текста вообще не слишком естественна.

Среди других корпусов стоит упомянуть созданный в Германии корпус Cosmas 1 (http://corpora.ids-mannheim.de/~cosmas/), включающий письменные, разговорные и диалектные тексты, некоторые из которых классифицированы по социальному положению авторов.

Помимо этого, на материале русского языка создан сборник диалектных текстов и текстов народов севера России [Герд и др. 2002]. Это собрание текстов, не являющееся корпусом в собственном смысле слова, было составлено на материале корпусных и диалектологических исследований, проводимых в Санкт-Петербургском государственном университете.

Спорные случаи в унификации

Хотя обычно отклонения, не связанные с морфологией, в текстах не отражаются, есть случаи, когда после изменений можно говорить уже об образовании новой лексемы. Такие примеры при разметке отражаются.

Так, литературному [жд] в диалектах соответствует [ж]: *Рожество*. Как и в случаях типа *робота*, грамматические показатели не меняются. С другой стороны, данный переход носит менее регулярный характер, чем оканье, хотя бы потому, что слов с последовательностью *жд* в русском языке не так уж много. Ещё дальше от обязательных переходов отстоят примеры типа *Паска* ‘Пасха’ — неверно, что в северных диалектах литературному [х] всегда соответствует [к]. Следовательно, слова типа *Рожество* и *Паска* с нерегулярными изменениями корня считаются отдельными лексемами и подаются с пометой *diallex* (возможность поиска по литературным словам сохраняется, поскольку после пометы указывается значение слова).

Проблемы морфологической разметки

Неоднозначность в приписывании морфологических категорий диалектным лексема может быть связана и с лексическими, и с морфологическими, и с синтаксическими особенностями диалектных текстов. В частности, морфологические особенности диалектов (переход большого количества прилагательных в класс на *-ой* создаёт большое количество новых случаев омонимии, о которой речь пойдёт ниже).

Среди чисто диалектных лексем неоднозначности в разметке часто возникают у глаголов, которым сложно приписать видовую характеристику. Ясно, что и в литературном языке однозначно приписать глаголу вид только исходя из его формы невозможно, но эта проблема даже усугубляется в диалектах: многие глаголы встречаются в тексте всего несколько раз в одной форме. Ср. примеры типа:

Как из ружья вверх ***стрелили***;

парень-от ишь ***запопивал***;

а осенью, и картошку выкопат всё, только ***сносят***.

В первом случае тип спряжения глагола отличается от литературного (*стреляли*), что не позволяет однозначно определить его видовую характеристику в диалектном тексте. Во втором примере задачу осложняет то, что глагол *запопивать* имеет две приставки и отличается от литературного. В третьем случае мы имеем дело с глаголом, который и в русском литературном языке является двувидовым.

Дополнительную трудность создают особенности употребления времён в северных диалектах: настоящее время в текстах часто используется в значении настоящего исторического, а прошедшее — в значении прошедшего хабитуального:

встанешь{вставать=V,непер=сов,изъяв,непрош,ед,2-л} *до*{до=PR} *солнышка*{солнышко=S,сред,неод=ед,род} *и*{и=CONJ} *под*{под=PR} *зорю*{зоря=S,жен,неод=ед,вин} *сходишь*{сходить=V,сов,непер=изъяв,непрош,ед,2-л}.

В результате оказывается, что настоящее и прошедшее употребляются в одних и тех же контекстах, следовательно, различение видов затрудняется. При невозможности сделать выбор, исходя из самой формы, например *парень-от ишь запопивал, они вот пьют друг у дружки* (контекст имеющего место состояния требует резу-

льтативной интерпретации формы *запопивать*, и ей приписывается совершенный вид).

Синтаксические *vs* морфологические особенности

В диалектах также имеются отклонения от литературной нормы из области синтаксиса: например, в первом собрании текстов местоимение *она* во всех косвенных падежах может иметь форму *ей*. Для таких случаев при разметке предполагается указывать, какую синтаксическую позицию занимает данная словоформа и какую падежную форму она бы имела в литературном языке (например, при употреблении формы *ей* в позиции прямого дополнения — *DO, ACC* — прямое дополнение, винительный падеж). Таким образом, случаи употребления одной формы в синтаксической позиции, характерной для другой, могут быть найдены как по одной форме, так и по другой: например, если исследователю нужно найти существительные и местоимения, выступающие в позиции прямого дополнения, то слово *ей* будет найдено как прямое дополнение, а если важно изучить употребление формы *ей*, то будут найдены все позиции, в которых она встречается. Правда, учесть при разметке все такие случаи не представляется возможным, потому что надёжных сведений о синтаксисе диалектных текстов описания не содержат. Например, встречается нестандартный эллипсис типа *Десятого числа было Самсону-Сеногною*, где неясно, к чему относится существительное в дативе. То же относится к предложным конструкциям вида *мы к Олонецкой губернии* (относимся, принадлежим).

В настоящее время синтаксические особенности диалектных текстов практически не отмечаются при разметке. С другой стороны, некоторые случаи являются спорными, допускающими как «морфологическую», так и синтаксическую трактовку.

Так, некоторые глаголы в диалектах имеют другое управление, чем в литературном языке: ср. *проведовать* ‘проведывать’: *Не ходи́ ко мне́ прове́довать, Я́ тебе бу́ду прове́довать*, где глагол, казалось бы, управляет дательным падежом. С другой стороны, в тексте присутствует только два вхождения данного глагола — в первом из них объект опущен, а во втором глагол управляет личным местоимением *тебе*.

Падежные формы местоимений в диалектах сильно отличаются от литературных: в частности, форма дательного падежа использу-

ется вместо формы винительного. У местоимения *она* смешение форм зашло особенно далеко — как было сказано выше, форма *ей* часто используется как универсальная для всех косвенных падежей. Следовательно, точно определить, что играет роль в приведённом выше примере — использование формы дательного падежа как формы винительного или изменение управления глагола — по приведённому отрывку нельзя.

Омонимия в диалектном корпусе

Поскольку диалекты во многом отличаются от литературной нормы, возникают случаи омонимии, отсутствующие или встречающиеся реже в литературном языке.

Прежде всего, это — омонимия форм прилагательных. Из-за распространённости в диалектах типа склонения прилагательных на *-ой* совпадают формы родительного падежа единственного числа женского рода и именительного падежа единственного числа мужского рода, в связи с чем возрастает количество вариантов разбора для каждого слова.

Также более распространена, по сравнению с литературным языком, омонимия кратких прилагательных и наречий. В диалекте вообще чаще, чем в литературном языке, встречаются стяжённые формы прилагательных. С этим также связана синтаксическая омонимия, поскольку не всегда легко установить, какую позицию — атрибутивную или предикативную — занимает имя.

Поскольку в диалектах имеются частицы, согласуемые по роду, числу и падежу, возникает омонимия между согласуемой и несогласуемой частицей *то*.

Устройство морфологической разметки

Общее устройство разметки слова в диалектном корпусе не отличается от принятой в литературных текстах: запись начинается с леммы, затем следуют словоклассифицирующие и словоизменительные характеристики. Отличается только четвёртое поле: если в «литературном» корпусе в это поле попадают различные нарушения, то в диалектном — специфически диалектные пометы.

В четвёртом поле морфологической разметки указываются диалектные характеристики слова. Прежде всего, даются пометы *dialmorph* — «диалектная морфология» и *diallex* — «диалектная лексема». Эти пометы позволяют задавать два различных типа поиска

— по диалектным лексемам (например, при составлении словарей) и по диалектным грамматическим явлениям (при составлении грамматических описаний).

Лемматизация

Лемматизация может проводиться как по русской литературной лексеме, так и по диалектной — это зависит от того, в чём заключается различие между диалектом и литературным языком. В случае, если различия заключаются в образовании данной формы, лемматизация проводится по литературной лексеме. Напротив, если мы постулируем диалектную лексему, лемматизация проводится по ней.

Так, в севернорусских диалектах встречается нулевой показатель третьего лица единственного числа настоящего времени: ср. *бу́дё* ‘будет’. В таких случаях нет оснований считать, что в диалекте используется нелитературная лексема: скорее нужно говорить об отличном от литературного грамматическом показателе. Именно поэтому в качестве леммы используется инфинитив *быть*, а не какой-либо другой. Даже если есть основания предполагать, что в исходной форме диалектного слова (например, в именительном падеже прилагательного) ударение стоит в другом месте, чем в литературном языке, в лемматизации это не отражается, потому что важно максимально облегчить поиск по литературным словам.

Такой способ лемматизации, безусловно, имеет свои недостатки. В частности, видя диалектную словоформу, мы можем только предполагать, что её исходная форма выглядит так же, как в литературном языке. К сожалению, объём диалектных тестов невелик, и каждая лексема встречается в немногих формах — следовательно, невозможно гарантировать, что инфинитив глагола ‘быть’ в диалектах выглядит как *быть*. В некоторых случаях точно определить исходную форму просто невозможно: например, неясно, какой инфинитив должна иметь форма *ложишь* — *класть* или *ложить* (не зная распределения вариантов, точно определить его нельзя).

Тем не менее, мы постулируем такое решение потому, что в сомнительных случаях всегда желательно приводить текст к литературному виду, чтобы облегчить поиск.

КОНКРЕТИЗАЦИЯ ПОМЕТ

Помета *dialmorph* конкретизируется: указывается, в чём состоит нарушение диалектной морфологии. Это делает более осмысленной характеристику слова: собственно характеристика формы как морфологически аномальной с точки зрения литературного языка не слишком информативна. С другой стороны, её конкретизация позволяет изучать диалекты с точки зрения конкретных грамматических проблем.

Виды помет:

- *flex* — слово имеет флексию, отсутствующую в парадигме данной части речи в литературном языке (*будё* — нулевая флексия в третьем лице единственного числа настоящего времени).
- *type* — слово имеет флексию, которая имеется в парадигме данной части речи, но в литературном языке эта лексема относится к другому словоизменительному классу (*хорошой* ‘хороший’). Наиболее распространено изменение типа прилагательных, реже встречаются отклонения у существительных (*на лошад**е*** ‘на лошади’). Ещё одной группой морфологических изменений, происходящих в диалекте, являются изменения типов спряжения (ср. *ревит* ‘ревёт’). При этом набор показателей, использующихся в диалектах, в основном совпадает с русским литературным, но их распределение по лексемам меняется, что затрудняет разметку и поиск. Таким образом, помета *type* встречается значительно чаще, чем *flex*.
- *part* — диалектная склоняемая частица (*-то/-от/-ту* и т. д.).
- *contr* — стяжённые формы прилагательных, совпадающие с краткими прилагательными в именительном падеже единственного числа, но изменяющиеся по роду, числу и падежу (*хорошу* ‘хорошую’).
- *refl* — глагол отличается от литературного только возвратностью (*гоститься* ‘гостить’).
- *stem* — при склонении не происходит изменений основы, характерных для данной лексемы в литературном языке (например, *пекёт* вместо *печёт* с отсутствием переходного смягчения).

К диалектной морфологии (*type*) относятся также отличные от литературных окончания неизменяемых слов, например, наречий (*когды*): хотя в корпусе этим словам не приписывается окончание, ясно, что изначально различия типа *когда*/*когды* восходят именно к различиям в окончании.

Помета *diallex* сопровождается коротким описанием значения слова. Это позволит проводить в диалектных текстах поиск по теме (например, «названия дома в диалектных текстах»).

Проблемы выбора пометы

Как было сказано выше, в корпусе диалектных текстов используются пометы *dialmorph* и *diallex*. В действительности не всегда ясно, какую именно помету следует выбрать в том или другом случае.

Морфологическая разметка предполагает, что в тексте встречаются практически все формы той или иной лексемы. Для диалектных текстов это неверно:

(1) нередко в текстах встречается только одна или меньшая часть форм от той или иной лексемы;

(2) часто в текстах встречаются формы с одной и той же грамматической характеристикой, различающиеся только типом склонения.

В приведённом ниже отрывке встречаются две формы творительного падежа единственного числа — *дедушкой* и *дедушком*:

*Па́па оста́л, у́мёр, нас че́тверо оста́лось, с **де́душкой** доживала́, а пото́м ба́бушкой. **Де́душка** у на́с бы́л жало́сливой-жало́сливой, серде́чной. Мне но́нь-то счита́ется, што я с **де́душком** пережила́, кака́я жи́знь была́ весе́лая, кака́я интере́сная.*

В данном случае существует два одинаково приемлемых решения:

1) приписать форме *дедушком* помету *dialmorph*, считая её формой от лексемы *дедушка*;

2) приписать форме *дедушком* помету *diallex*, считая её формой от диалектной лексемы *дедушко*. Склонение морфологически уменьшительных существительных по типу *озеро*, как существительных среднего рода, в диалектах очень распространено: ср. *хозяинушко*, *батюшко*, *Леконидушко*. С другой стороны, постулировать данный вариант как единственный невозможно из-за наличие формы *дедушка*.

Первое решение можно назвать *морфологической стратегией*: в данном случае словарь остаётся таким же, как и в диалектах, зато максимально расширяются грамматические возможности. Второе решение — проявление *лексической стратегии*: при сохранении грамматических правил постулируются новые лексемы.

В действительности предпочтительнее выглядит второе решение: при наличии отрывочных сведений о диалектах постулировать новые правила грамматики явно не стоит.

Другая сложность в выборе помет связана с изменяемой постпозитивной частицей: ср. *полуночница-**та***, *день-**от***, *по дороге-**то***. По своим функциям она схожа с русской частицей *то* в конструкциях типа *Он-то точно это знает*, а по формальным свойствам — с постпозитивными артиклями балканских языков.

Проблема в данном случае заключается в том, (1) происходят ли при присоединении данной частицы грамматические процессы, отличные от литературных, и (2) считать частицу диалектной лексемой или приписывать ей диалектную морфологию (согласование с характеристиками управляющего слова).

Заметим, что выбор формы частицы можно связать не только с грамматической характеристикой имени: в частности, большую роль играет консонантное/вокалическое окончание имени: ср. *вредны-**ти*** и др. В частности, частица в формах типа *слушать-**от*** имеет форму, характерную для слов с консонантным окончанием, но не обусловлена грамматической характеристикой слова. Следовательно, можно было бы считать, что грамматических процессов, отличных от литературных, в данном случае не происходит, а форма частицы изменяется по особым фонологическим правилам (ср. *пошёл бы — *пошёл б — пошла б* и т. д.). В таком случае данные изменения снабжались бы пометой *dialmorph* на тех же основаниях, что и фонетически мотивированные изменения в типе склонения.

С другой стороны, отмеченные выше различия между формами с одинаковыми внешне окончаниями, но разной грамматической характеристикой мешают считать изменение частиц чисто фонетическим процессом.

Второй вопрос связан с тем, как можно проводить лемматизацию частицы. В случае, если частицам будет приписываться часть речи PART, необходимо постулировать и диалектную лексему, и диалектную морфологию, поскольку в литературном русском языке частицы не изменяются по родам, числам и падежам. Более экономичным было бы приписать частице часть речи A (прилагательное), не отличая её от местоимения *тот*.

Тем не менее, с грамматической точки зрения было бы более обоснованным отнести частицу к той же лемме, что и частицу *то* (исходной формой считается именно *то*, а не *тот*, что подчёрки-

вает близость изменяемых и неизменяемых частиц, часто сосуществующих в одном и том же тексте). При частице даётся помета *dialmorph*, поскольку в действительности изменяемые и неизменяемые частицы не являются разными лексемами.

Ещё одну проблему составляют словоформы, отличающиеся от литературных не словоизменительными характеристиками, а словообразовательными формантами, в случаях, когда параллель с литературным языком достаточно явная: ср. *взамуж* ‘замуж’, *зауснёшь* ‘уснёшь’, *проведовать* ‘проведывать’ и др. Подобные словоформы не являются формами литературных лексем и, говоря формально, должны помечаться как диалектные лексемы.

С другой стороны, полезно и указание на диалектную морфологию, хотя разные лексемы в разной степени могут считаться морфологическими вариантами литературных. В частности, формы типа *взамуж*, возможно, вызваны переосмыслением лексемы *замуж* как простого слова, а следовательно, представляют собой следующий этап словообразования, отстоят довольно далеко от образования лексемы *замуж* и с трудом могут считаться её морфологическим вариантом, поскольку литературное и диалектное слова образованы от разных основ. Напротив, лексема *проведовать* имеет такой вид в результате переосмысления суффикса в литературной лексеме *проведывать*, и это сближает её с вариантами типа *дедушком/дедушкой*, где исходная основа одна и та же, но варианты словоизменения и словообразования различны. Кроме того, видовые суффиксы близки к словоизменительным аффиксам. Особый случай представляют лексемы, отличающиеся от литературных только возвратностью или её отсутствием (*дождать* ‘дождаться’, *гоститься* ‘гостить’).

Также существуют словоформы и лексемы, не являющиеся диалектными, но принадлежащие к разговорному языку и поэтому отсутствующие в большинстве диалектных словарей: например, частицы *дак*, *от* и т. д. Поскольку в записях литературных текстов они обычно отсутствуют, в диалектном корпусе они также помечаются пометой *diallex*. С другой стороны, некоторые частицы принадлежат только диалектам — например, *ак* ‘а’.

Эллипсис

Ещё одна проблема связана с эллиптичностью устного текста, которая не всегда позволяет выявить грамматическую характери-

стику словоформы: ср., например, *Звёзды е́сь на не́бе. Примеча́ют, звёзды пока́зывают, забы́л, како́й приме́т*, где возможна и интерпретация *звёзды* (номинатив) *показывают* (*наше будущее*), и (*люди*) *показывают звёзды* (аккузатив). Более того, даже предложения, интерпретация которых сомнений не вызывает, часто представляют трудности с точки зрения синтаксического разбора — ср., например, *ох уж раньше двадцать-то лет*, со значением ‘двадцать лет мне было давно, а теперь уже гораздо больше’. Данное предложение находится в тексте после предложения *как в двадцать лет*, а значит, не вполне ясно, в каком падеже стоит слово *двадцать*. Как правило, в подобных случаях из возможных прочтений выбирается такое, которое наиболее близко к словарной форме слова или совпадает с ней.

Членение текстов

Как для изучения синтаксиса диалектов, так и для облегчения их восприятия необходимо обозначать в корпусе членение текстов на предложения. При этом возможны два решения: с одной стороны, паузы различных типов могут обозначаться знаками препинания: это придаёт текстам вид, более привычный для русского читателя. Однако недостаток этого подхода в том, что сами знаки препинания часто отражают синтаксическую, семантическую и прагматическую структуру текста очень сложным образом, во всяком случае, не прямо, а опосредованно.

Для научных задач более приемлема запись с фиксацией длины пауз, интонации и, возможно, других характеристик текста. Однако такая запись может затруднить чтение текстов и поиск в них. Кроме того, не все исследователи отмечают паузы во всех предложениях — если изучается грамматика, а не интонация или фонетика, используется обычная система знаков препинания. В частности, в нашем корпусе только в материалах второй и третьей экспедиции отмечаются паузы и маркируется их длина (как можно заметить, фонетически и интонационные характеристики речи вообще отмечаются более тщательно именно во второй и третьей подборке текстов). В материалах первой экспедиции используется только стандартная пунктуация. Поскольку восстановить паузы в данных текстах не представляется возможным, во всех диалектных текстах будут использованы стандартные знаки препинания, а во второй и третьей подборке — также знаки пауз.

С помощью особого формата выделяются реплики исследователей, не являющиеся частью текста на диалекте.

Поскольку в один текст объединяются реплики разных информантов на одну и ту же тему, они различаются с помощью деления на абзацы — один абзац всегда принадлежит только одному информанту.

Деление корпуса

Корпус диалектных текстов делится на подкорпусы, а они, в свою очередь, на отдельные тексты.

Подкорпусы выделяются по принадлежности текстов к той или иной системе говоров: в частности, в нашем материале есть тексты из южных и из северных говоров. Три северных говора в нашем материале не только различаются между собой, в том числе и по грамматике, но и изучены с разной степенью подробности: например, только для первого из них специально указаны грамматические особенности, для остальных указываются почти исключительно фонетические. Кроме того, подкорпусы существенно различаются по метаразметке. Тем не менее они должны входить в один подкорпус — различия в их грамматике должны указываться в комментариях к каждому из текстов.

Деление подкорпусов на тексты не всегда совпадает с выделением отрывков в экспедиционных материалах, но часто соответствует ему. Тематическое деление сохраняется: как правило, авторы материалов первого говора разбивали тексты на довольно маленькие отрывки, каждый из которых посвящён одной теме. Авторы второго и третьего корпусов практически не членили тексты по темам.

Могут возникнуть случаи, когда текст сложно разделить на части по темам — например, некоторые из частей оказываются незавершёнными, поскольку носитель может переходить от теме к теме. В этом случае в поле «тема» может указываться несколько вариантов, а части могут объединяться. Так, три предложения: *27. А ещо́ ка́шей называ́ли, я тебе́ говори́ла, пожина́ха. И ребёнка крести́ли — ка́ша. 28. А крести́ли ребя́т да всё куми́лися. Все́ у Дю́ковой избы́ забрели́ в во́ду, да во́т о́н и крести́л в реке́. 29. Ско́лько пи́сен-то стари́ных бы́ло пи́то. А каг запою́т-то — по ле́су гу́т стои́т* — исследователи, составлявшие подборку текстов, включили в часть «Кален-

дарь. Праздники». В корпусе они также будут отнесены к этой теме. Хотя у этих трёх предложений разные авторы, разделять их нет особого смысла.

Часто бывает разумно разделить тексты менее дробно, чем это сделано диалектологами. В подборке текстов деление неравномерно, например, объединены все тексты по теме «Календарь. Праздники», но разделены «Вода. Водяной» и «Лес. Леший». По всей вероятности, имеет смысл объединить две последних темы в тему «Сверхъестественные существа». Крупное деление имеет смысл ещё и потому, что тексты второго говора разделить в целом сложнее, темы там выделяются не столь явно, как в первой подборке.

Метаразметка

Метаразметка диалектных текстов почти идентична принятой для корпуса устных текстов. Тем не менее, есть определённые отличия и от литературной, и от «устной» метаразметки.

Автор текста, как правило, не указывается. Обычно диалектологические экспедиции собирают тексты от разных носителей, каждый из которых записывает очень небольшой текст. Исследователи авторами не считаются — их реплики в текст не попадают или почти не попадают (это зависит и от способа записи текстов). В этом смысле метаразметка диалектных текстов похожа на представление в корпусе интервью, где реплики журналиста, во-первых, занимают меньшую часть текста, а во-вторых, менее важны, чем реплики интервьюируемого. Следовательно, все диалектные тексты являются монологами, хотя имеют некоторые черты, характерные для диалогического общения (в частности, содержат обращения к адресату типа *я ж тебе говорила* и т. д., предполагающие его реакцию). Вопросы исследователя включаются в текст без грамматического разбора, поскольку не являются текстом на диалекте.

В то же время некоторые сведения об авторе необходимы для разметки: прежде всего, пол, возраст и происхождение (если эта информация доступна). Особенно важен возраст автора, поскольку население многих деревень состоит почти исключительно из пожилых людей, и диалект не имеет молодых носителей.

Набор тем текстов мало отличается от литературного, но, естественно, гораздо более ограничен. Поскольку диалектные тексты посвящены почти исключительно быту и обычаям, необходимо

более точно указывать тему (например, «Сверхъестественные существа», «Обряды» и т. д.). Названия текстов представляют собой уточнённую информацию о теме: например, если в теме указывается «праздники», текст может называться «Празднование Крещения» и т. д. Точное представление темы текста удобно для пользователей, использующих корпус для социолингвистических целей: например, для сравнения представлений о нечистой силе в русских диалектах.

Тексты помечаются как публичные, поскольку произносятся специально для исследователя по его просьбе и для записи. Образцы спонтанной речи в диалектных материалах не встречаются.

Все тексты попадают в класс «рассказ»; возможно, очень небольшая часть будет отнесена к классу интервью. Отделяются от основного текста стихи и песни, так как, судя по всему, их авторство не принадлежит автору текста. Кроме того, не вполне ясно, отличается ли грамматика стихов и песен от основного текста.

В поле «комментарий» записывается краткая информация об особенностях диалекта, прежде всего, фонетических и морфологических. К сожалению, корпуса прокомментированы с разной степенью подробности. Также указываются исследователи, собиравшие материалы, поскольку в поле «автор» они, как было сказано выше, не попадают. Кроме того, в поле «комментарий» записывается информация о значении диалектных лексем, которую невозможно вместить в разметку каждого слова.

Поскольку помещенные в Корпус диалектные тексты были опубликованы, заполняются поля «Тип издания», «Место издания» и «Время публикации», но кроме этого необходимо указывать место и время произнесения самого текста. Для этого существуют отдельные поля метаразметки. Помимо обычного для метаразметки указания места записи, указывается группа диалектов, к которым относится данный говор.

Заключение

Как я попытался показать, набор грамматических категорий и частей речи для разметки диалектного корпуса не отличается от литературного — за исключением того, что некоторые части речи, например, частицы и прилагательные в краткой форме могут иметь категории, которых не имеют в литературном тексте. Мета-

разметка диалектных текстов сильно отличается от литературной, но похожа на разметку устных текстов.

Наконец, есть аспекты, в которых диалектный корпус отличается и от литературного, и от «устного» — это специфические пометы в последней части морфологической разметки, позволяющие идентифицировать слово или его грамматические характеристики как диалектные.

Проблемы в разметке диалектных текстов касаются именно этих специфических характеристик, и, как следствие, также лемматизации (лемматизация связана с тем, считаем ли мы нужным постулировать диалектную лексему). Во многом эти проблемы связаны с общими особенностями диалектных текстов: (1) небольшим объёмом, (2) сильной вариативностью и (3) отличиями как от русского литературного языка, так и от устной речи.

Приведем несколько примеров разметки:

Словоформа с фонетическим переходом, затрагивающим морфологические показатели: *кумитсе* [кумиться= V несов, непер, мед = инфс = dialmorph, refl]

Словоформа с диалектным показателем: *будё* [быть= V, непер= несов, изъяв, непрош, ед, 3-л = **dialmorph.**, flex]

Диалектная лексема: *пожинаху* [пожинаха =S, жен, неод=ед,вин = **diallex** ,'каша']

Диалектная лексема, имеющая литературное соответствие с тем же корнем: *взамуж* [взамуж (замуж)=Adv=diallex,'замуж']

Диалектная лексема, состоящая из литературных слов: *в-под* [в-под (в+под)= PR = diallex]

Литература

Гришина Е. А. Устные тексты в корпусе русского языка. // НТИ, 2005, № 3.

Герд А. С., Савиярви М., де Грааф Т. (ред.). Язык и народ. Тексты и комментарии. СПб.: СПбГУ, 2002.

Летучий А. Б. Диалектные тексты в корпусе русского языка.// НТИ, 2005, № 3.

Описание корпуса Cosmas 1. http://corpora.ids-mannheim.de/~cosmas/.

Peitsara K., Vasko A.-L. The Helsinki Dialect Corpus: characteristics of speech and aspects of variation. http://www.eng.helsinki.fi/hes/Corpora/helsinki_dialect_corpus.htm. 2002.

Qian Yue Liang, Lin Shou Xun, Zhang Yong Dong, Liu Yang, Liu Hong, Liu Qun. An Introduction to Corpora Resources of 863 Program for Chinese Language Processing and Human-Machine Interaction. http://mtgroup.ict.ac.cn/~liuyang/papers/final_revised.pdf. 2000