КОРПУС РУССКОЙ ДИАЛЕКТНОЙ РЕЧИ: КОНЦЕПЦИЯ И ПАРАМЕТРЫ ОЦЕНКИ[1]

О. Ю. Крючкова (vpks@rambler.ru)

В. Е. Гольдин (goldinve@yandex.ru)

Саратовский государственный университет
им. Н. Г. Чернышевского, Саратов, Россия

На основе сравнительной оценки двух диалектологических корпусов — диалектного корпуса в составе Национального корпуса русского языка (ДК НКРЯ) и Саратовского диалектологического корпуса (СарДК) — обсуждаются концепция и параметры оценки корпуса русской диалектной речи.

Ключевые слова: диалект, диалектный корпус, диалектная речь, параметры.

A CORPUS OF RUSSIAN DIALECTAL SPEECH: THE CONCEPT AND PARAMETERS OF EVALUATION

O. Iu. Kriuchkova (vpks@rambler.ru)

V. E. Gol'din (vpks@rambler.ru)

Saratov State University, Saratov, Russian Federation

The concept and parameters of evaluation of a corpus of Russian dialectal speech are discussed based on the comparative assessment of two dialect corpora — dialect corpus within the National Corpus of the Russian Language (DC NCRL) and the Saratov Dialect Corpus (SarDC): the principles of selection of the dialect materials and the criteria of the dialect corpus representativeness; the principles of the speech continuum partition in the corpus; the parameters of textual fragments return; the forms of representation of the dialect texts in the corpus; the types and rules of annotation of the corpus textual basis; the parameters of the dialect texts meta-marking; the representation of nonlinguistic information in the corpus; the possibilities of retrieval queries, optimal for dialect research. The paper proves that the dialect corpus cannot be based on the same model as the corpus

[1] Работа выполнена при поддержке Фонда «Русский мир» (грант № 354 Гр/1232-10).

of standard language because of the specifi c character of the dialect material. The dialect corpus must be modeled as a system of corpora of different dialects, representing the main dialect types of the Russian speech. According to the proportionality principle, the textual basis of the corpus of a separate dialect must be aimed at the modeling of communication in this specifi c dialect, refl ecting the main types and forms of the dialect speech, as well as social differentiation of the dialect native speakers and genre and theme structure of the dialect communication.

Key words: dialect, dialectal corpus, dialectal speech, parameters.

1. Введение

Сделанная в Национальном корпусе русского языка (НКРЯ) попытка создать на его лингвистической и программной платформе диалектный корпус русского языка, с одной стороны, продемонстрировала несомненную ценность применения к диалектному материалу общих для НКРЯ классификационных и структурных принципов; с другой стороны, результаты этой попытки и анализ параллельного развития других диалектных корпусов (например, Саратовского диалектологического корпуса) высветили целый ряд проблем, без решения которых корпусная диалектология и общерусский диалектный корпус не могут эффективно развиваться. Сейчас, когда диалектологами и разработчиками НКРЯ осознано неудовлетворительное состояние ДК НКРЯ, когда «пришло понимание, что этот проект полезно было бы перестроить так, чтобы он служил самим диалектологам — и как удобно организованный ресурс для учебного процесса, и как инструмент для исследовательской деятельности» [Рахилина 2009: 15], существует острая потребность в новых концептуальных решениях.

Необходимы выработка четкой концепции корпуса, базирующейся на выделении системы важнейших его параметров, и принятие в соответствии с каждым из них определенных решений. К числу таких параметров относятся, по нашему мнению, следующие:

1. принципы отбора диалектного материала и критерии репрезентативности диалектного корпуса;
2. принципы членения речевого континуума в корпусе;
3. параметры выдачи текстовых фрагментов;
4. формы представления диалектных текстов в корпусе;
5. виды и правила аннотирования текстовой базы корпуса;
6. параметры метаразметки диалектных текстов;
7. представление в диалектном корпусе нелингвистической информации;
8. оптимальные для диалектологических исследований возможности пользовательских запросов.

Дальнейшее изложение посвящено обсуждению названных параметров, которое проводится на основе сравнительной оценки двух диалектологических

корпусов — диалектного корпуса в составе Национального корпуса русского языка (ДК НКРЯ) и Саратовского диалектологического корпуса (СарДК[2]).

2. Параметры оценки диалектного корпуса

1. Принципы отбора диалектного материала и критерии репрезентативности диалектного корпуса

Необходимым условием решения данного вопроса является определение ***цели*** диалектного корпуса. Он может создаваться либо как корпус иллюстративного типа, единственная цель которого — демонстрация территориальной неоднородности национального языка, либо как научный источник нового типа, соответствующий общей идеологии корпусной лингвистики.

В первом случае репрезентативность диалектного корпуса определяется прежде всего максимальным охватом территорий распространения национального языка; при этом допустимы фрагментарность материала и отсутствие системности его представления. Такой корпус носит ознакомительный, популяризаторский характер, но не имеет ценности в качестве источника лингвистических или лингвокультурологических исследований[3]. Подобной модели диалектного корпуса соответствует сегодня ДК НКРЯ: «Корпус проектировался и создавался, — пишет Е.В. Рахилина, — с ориентацией на… рядовых пользователей…, большинство из которых никогда в жизни не видело ни одного диалектного текста. В то время задачей было сделать своего рода «научную игрушку», которая наглядно демонстрировала бы разнообразие русского языка в его региональных вариантах» [Рахилина 2009: 15]. ДК НКРЯ собирает сегодня по-разному записанные и по-разному же обработанные какие угодно текстовые фрагменты любых русских говоров и территориально соотносит их с областными центрами или с еще более крупными географическими ориентирами (*Архангельск, Вологда, Курск, Саратов, Забайкалье, Карелия*). Такой материал при любом увеличении его объема не может стать репрезентативным представлением ни отдельных русских говоров как особых полносистемных образований, ни русской диалектной речи в целом.

Диалектный текстовый корпус не может строиться по той же модели, что и корпус «стандартного» (литературного) языка ввиду специфичности диалектного материала, учет которой необходим при создании адекватных природе говоров и по-настоящему репрезентативных диалектных корпусов. Представим часть из этих различий в табличной форме (см. Табл. 1)

---

[2] Саратовский диалектологический корпус разрабатывается в Центре изучения народно-речевой культуры им. профессора Л.И. Баранниковой Института филологии и журналистики Саратовского государственного университета им. Н.Г. Чернышевского.

[3] Именно так строятся диалектологические корпуса многих других языков.

Таблица 1. Различия между текстами «стандартного языка» и текстами диалектными

Тексты «стандартного» языка	**Диалектные тексты**
1. «Стандартные тексты» (прежде всего письменные) в основном закреплены в составе каких-то изданий, собраний, библиотек, фонотек и т. п., в этом виде они естественным образом функционируют в преимущественно городской коммуникации и доступны наблюдателям.	1. Диалектные тексты не закреплены в их естественной (традиционной сельской) устной коммуникации, они получают закрепление и становятся доступными наблюдателям лишь в составе внеположенных диалектной коммуникации диалектных корпусов, хрестоматий и др. научных источников.
2. Достижение пропорциональности и репрезентативности корпуса русской литературной речи становится всё более простым и быстрым благодаря распространению электронных форм воплощения текстов.	2. Вследствие исключительно устного воплощения диалектной речи и множества относительно самостоятельных русских говоров при специфичности их внутренней организации репрезентативность национального корпуса по отношению ко всему «русскому диалектному языку» в составе его «микросистем» реально не может быть в обозримое время достигнута, тогда как репрезентативность материалов по отдельным давно и подробно изучаемым говорам, хорошо отражающим его главные структурные части (наречия, группы, зоны), вполне достижима уже сегодня при условии трансформации собранных диалектологами материалов в корпусную форму и целенаправленном их пополнении.
3. «Стандартные тексты» подготовлены самими авторами или публикаторами к использованию в публичном общении, рассчитаны на это, и, следовательно, открытие их в корпусе не является неразрешенным вторжением в чью-либо личную сферу.	3.Диалектные тексты в основном имеют более личный, часто даже глубоко интимный и просто наивно-открытый, незащищенный характер и были бы совершенно другими (или вовсе не состоялись бы), если бы говорящие предполагали, что их речь будет вынесена на всеобщее обозрение.

Тексты «стандартного» языка	Диалектные тексты
4. «Стандартные тексты» являются частью той культуры (в том числе языковой), к представителям которой относятся создатели корпуса и предполагаемые пользователи, поэтому тексты относительно легко могут подготавливаться к включению в корпус любыми филологически грамотными людьми и/или специальными компьютерными программами. По той же причине содержание текстов (упоминаемые в «стандартных текстах» события, лица, природные объекты, артефакты, идеи и т. п.) в большинстве случаев не требуют специального комментирования.	Диалектные тексты представляют совершенно особую, так называемую «традиционную культуру», особые самодостаточные языковые системы и автономные коммуникативные образования. Они воплощают специфическое содержание и специфические формы коммуникации, поэтому без специального лингвистического и культурологического сопровождения эти тексты могут лишь казаться понятными предполагаемым пользователям корпуса.

Следствиями отмеченных различий между «стандартными» текстами и текстами диалектными являются следующие общие принципы отбора и корпусного представления материала, без соблюдения которых репрезентативность диалектного корпуса как научного источника не может быть достигнута:

1.1. Диалектный корпус должен делать доступными по запросам не только фрагменты текстов, но и целые тексты, то есть диалектный корпус, в отличие от корпуса литературного языка, должен быть одновременно и диалектной библиотекой или архивным собранием материалов.

1.2. Диалектный корпус, как и обычные архивы, должен предусматривать различные степени допуска к различным его материалам (одну степень — для составителей, исследователей, другую для любых желающих) и постепенно открывать текстовые фонды, когда это становится возможным.

1.3.1. Диалектные тексты не могут адекватно пониматься, будучи вырванными из общего контекста родной для них традиционной культуры, поэтому они требуют воссоздания в корпусе соответствующего лингвистического и культурного (в самом широком смысле) фона в целом и в связи с содержанием каждого конкретного текста в отдельности. Эта проблема решается отсылками к размещенным в корпусе историческим, этнографическим, географическим и др. энциклопедическим данным мультимедийного характера, а также специальными комментариями к упоминаемым в конкретных текстах событиям, лицам, природным объектам, артефактам, идеям и т. п. Подобные комментарии целесообразно ориентировать не на традиционную культуру в целом и не на диалекты вообще, а на комплексы текстов конкретных говоров.

1.3.2. Подготовка диалектных текстов к введению в корпус — процесс более сложный и трудоемкий, чем подготовка «стандартных» текстов.

Он включает установление контакта с диалектоносителями, организацию записи, расшифровку, значительную долю ручной разметки, необходимо дополняющей автоматическое аннотирование, семантический и грамматический анализ. Это не механическая, а исследовательская работа, серьезный и очень ответственный авторский труд. Он может выполняться только специалистом-диалектологом, профессионально изучающим конкретный говор.

1.4. Диалектный корпус должен строиться как система корпусов отдельных говоров, представляющих важнейшие диалектные типы (наречия, группы, зоны) русской речи. В соответствии с принципом пропорциональности текстовая база корпуса отдельного говора должна стремиться к моделированию коммуникации в конкретном говоре, отражая важнейшие типы и формы диалектной речи, социальную дифференциацию носителей говора, жанрово-тематическую структуру диалектного общения.

Следование этим принципам обеспечивает репрезентативность диалектологического корпуса как научно-исследовательского источника.

2. Членение речевого континуума в диалектном корпусе

В соответствии с различными целями диалектологических корпусов в них по-разному решается задача членения представляемой в корпусе речи.

Текстовая база ДК НКРЯ наполняется отрезками диалектной речи (как правило, небольшого объема — в среднем не более 1 тыс. словоупотреблений), предварительно (до включения в корпус) фрагментированными на тематической основе. Членение речевого потока в СарДК отвечает принципу максимального приближения модели к объекту — естественной коммуникации на диалекте. Текст в СарДК полностью соответствует зафиксированному аудио- или видеоаппаратурой участку непрерывного общения, поэтому границы текста не зависят от таких параметров, как смена темы, жанра, формы речи, частичное изменение коммуникативной ситуации и числа ее участников. Такое представление речи существенно расширяет возможности использования корпуса.

3. Параметры выдачи текстовых фрагментов

В ДК НКРЯ в настоящее время предусмотрены 2 возможности выдачи: минимальный контекст и его расширение (как правило, до 3–4 строк, хотя тип расширения пока не носит последовательного характера). Однако специфика диалектного материала требует контекстов большей протяженности и возможности получения целой записи, поэтому в СарДК минимальной выдачей является абзац, а максимальной — целый текст, обычно значительной протяженности.

4. Формы представления текстов в диалектном корпусе

В ДК НКРЯ диалектные тексты представлены только в виде полуорфографической записи. Такая фиксация диалектной речи не позволяет изучать ее фонетическую сторону, что вызывает обсуждение вопроса о необходимости параллельного представления в ДК НКРЯ фонетической транскрипции. Однако в значительных по объему текстовых корпусах, в наполнении которых принимают участие большие и часто разрозненные коллективы диалектологов, достичь единообразия при транскрибировании весьма различных

по фонетической структуре диалектных текстов практически невозможно. В этих условиях бóльшую актуальность приобретает вопрос о включении в корпус аудио- и видеозаписей диалектной коммуникации и формах их соотнесения с символьной расшифровкой. В НКРЯ уже имеется успешный опыт создания устного корпуса со звуковой составляющей [см.: Гришина 2009], который важно использовать и при разработке диалектного корпуса. Отрадно, что такая перспектива уже заявлена [см.: Рахилина 2009: 14].

В СарДК параллельное представление текстовых и аудио-/видеомодулей является одним из важнейших принципов его строения, обеспечивающим максимальную достоверность информации. Наличие в корпусе звукового компонента обусловливает достаточность полуорфографической расшифровки диалектных текстов. Использование этой формы символьного представления речи в свою очередь требует решения вопроса о необходимости единого формата или достаточной степени единообразия полуорфографического представления диалектного текста в корпусе. В настоящее время в ДК НКРЯ нет единообразия символьного представления записей: расшифровки выполнены в разных диалектологических центрах по разным правилам. В СарДК используется единый формат символьной записи, создана специальная инструкция, регламентирующая характер отражения в расшифровке диалектных особенностей, способ членения текста и использование знаков препинания, способы обозначения нераспознанных фрагментов речи и недоговоренных слов, способы дифференциации речевых отрезков, принадлежащих диалектологу и диалектоносителю, способы дифференциации речевых отрезков, принадлежащих разным диалектоносителям, способ подачи необходимых для понимания текста комментариев.

5. Виды и правила аннотирования текстовой базы корпуса

Цель создания корпуса и характер включаемых в него текстов определяют принятые в нем виды и правила аннотирования текстовой базы. ДК НКРЯ и СарДК различаются как применяемыми видами разметки, так и правилами аннотирования при использовании одного и того же вида разметки.

Основным видом разметки в ДК НКРЯ и в СарДК, как и в большинстве текстовых корпусов, является морфологическая разметка, при проведении которой между сопоставляемыми корпусами есть существенные различия. Иллюстративная по своей сути стратегия ДК НКРЯ обусловливает использование дифференциального подхода при морфологической разметке корпуса. Диалектная речь представляется в ДК НКРЯ через ее соотнесение с литературной, рассматривается как речевая среда, характеризуемая отклонениями от литературных форм. Диалектные формы характеризуются такими, например, параметрами, как «другая флексия», «нестандартная флексия». Так, в качестве примеров сущ. Им. п. мн. ч. с «другой флексией» выдаются «деулинские» *окунья* и *утяты* или «волгоградское» *соседы*. «Другими» или «нестандартными» приведенные формы являются бесспорно лишь по отношению к литературной норме, но могут быть вполне системными для соответствующих говоров. Однако вопрос о системности конкретных говоров вообще не может ставиться и решаться на материале диалектного корпуса, если он строится по модели, которую можно было бы охарактеризовать как иллюстративно-дифференциальную.

В ДК НКРЯ используется сложная система дифференциальных помет, описывающая типы «отклонений» от литературной нормы [см.: Летучий 2005, 2009]. Стремление к детальной характеризации диалектных особенностей не только противоречит общему принципу НКРЯ «не навязывать пользователю своих исследовательских решений», но и в большей мере, чем для других подкорпусов НКРЯ, создает опасность навязывания субъективных квалификаций. Диалект, в отличие от других нелитературных разновидностей национального языка, — полносистемное образование, грамматическая специфика которого, как и специфика любого самостоятельного языка, не может быть описана без специального научного изучения репрезентативного языкового материала. Такое изучение и такое описание возможны лишь после создания полноценного диалектного корпуса конкретного говора, но никак не до того, как репрезентативность корпуса будет достигнута. Неизбежные уточнения и изменения в системе дифференциальных помет, возникающие по мере накопления материала (ср., напр. [Летучий 2005] и [Летучий 2009]), увеличивая трудоемкость процесса создания корпуса, не решают проблемы достоверности разметки.

СарДК в отличие от ДК НКРЯ — корпус принципиально недифференциального и нелитературноцентрического типа. Этим обусловлен ряд его отличий СарДК от ДК НКРЯ в лексико-морфологической разметке текстов (см. [Крючкова 2007]).

В связи с морфологическим аннотированием диалектного корпуса своего решения ждут также вопросы о лемматизации диалектных словоформ, о целесообразности и способах выделения в разметке различных функционально маркированных элементов в диалектной речи (просторечных, архаических форм); о целесообразности специальной маркировки различных видов неоднословных и идиоматических единиц. В ДК НКРЯ позиция по этим вопросам строго не эксплицирована. В СарДК приняты соответствующие (хотя, возможно, и не окончательные) решения и действуют определенные правила разметки названных единиц.

ДК НКРЯ и СарДК различаются по характеру тематической и жанровой разметки корпусов. В ДК НКРЯ выделение текста по тематическому принципу делает избыточной тематическую разметку каждого отдельного текста. Незначительный объем тематически цельных текстов в ДК НКРЯ обусловливает также нецелесообразность их жанровой разметки: каждый текст обычно оказывается моножанровым («бытовая сфера»).

В СарДК значительный объем, политематичность и полижанровость текстов, напротив, обусловливают необходимость такой разметки. Тематическая и жанровая разметка значительного по объему корпуса диалектных текстов в перспективе даст возможность исследовать жанрово-тематическую структуру диалектной коммуникации, выявить соотношение различных тем и жанров в составе диалектной коммуникации.

6. Параметры метаразметки диалектных текстов

В ДК НКРЯ метаописание текста ограничивается указанием на место, время записи, эксплоратора, общую тему и объем текста. Ввиду указанной выше специфики диалектных текстов названных параметров явно недостаточно для понимания и анализа диалектной коммуникации. Важными оказываются сведения о конкретной ситуации записи текста (в доме информанта,

в поле, в огороде, в лесу и т. п.), об адресатах речи, об упоминаемых в тексте лицах, о времени описываемых в тексте событий, о жанрово-тематической структуре записанного фрагмента речи. Все эти сведения включены в метаразметку текстов в СарДК.

7. Представление в диалектном корпусе нелингвистической информации

В ДК НКРЯ нелингвистическая информация ограничена приведенными параметрами метаописания. В СарДК, который создается как модель традиционной сельской коммуникации на диалекте, отражающая речевое общение в конкретных условиях жизни конкретного речевого коллектива, нелингвистической информации отводится специальное место. База корпуса включает отдельные модули с нелингвистической информацией, часть из которых связана с конкретными текстами, другая часть такой привязки не имеет. Текстовую привязку имеют биографический (биографические сведения об информанте) и иллюстративный (фотографии информанта, фотоиллюстрации к данному тексту) модули. Самостоятельный блок информации (не связанный с конкретным текстом) образуют другие справочные материалы: сведения исторического, социокультурного характера, данные демографические, этнографические, географические.

8. Оптимальные для диалектологических исследований возможности пользовательских запросов

Оптимальной для диалектологических исследований будет система пользовательских запросов, удовлетворяющая как потребностям уровневых описаний диалектной речи, так и исследовательским потребностям в области коммуникативного, лингвокогнитивного и лингвокультурологического изучения диалектного общения.

Применение к диалектному материалу общей для НКРЯ детально разработанной и разветвленной системы поисковых запросов является, безусловно, сильной стороной ДК НКРЯ и теоретически открывает перед диалектологами новые перспективы в изучении диалектной коммуникации. Однако эти возможности существенно ограничены (или даже сводятся к нулю) принципиальной нерепрезентативностью корпуса, обусловленной отсутствием концепции корпуса, соответствующей диалектному материалу.

СарДК предоставляет пользователю комплексную информацию о каждом конкретном говоре. Система пользовательских запросов в СарДК позволяет составлять выборки по отдельным морфологическим и лексическим явлениям, по тематическому и жанровому критериям, по отдельному информанту, по отдельному подкорпусу (говору) или по всем включенным в корпус подкорпусам. От текстовых модулей возможен переход к звуковым модулям и наоборот, а также параллельное их воспроизведение.

3. Заключение

Разработка диалектологических корпусов в настоящее время находится на начальной стадии, постоянно уточняются общие принципы и частные

методики их построения, поэтому широкое обсуждение обозначенных в данной статье проблем является актуальной задачей корпусной диалектологии и залогом ее успешного развития.

При условии выработки общей концепции русского диалектного корпуса, отвечающей специфике диалектного материала и современным задачам диалектологии, создания четких инструкций по всем параметрам обработки диалектного материала для включения в корпус, организации постоянно действующих семинаров для диалектологов, участвующих в наполнении корпуса, можно безусловно надеяться на то, что стадия «мечтаний, споров, проб... ошибок», в которой сейчас находится разработка диалектного корпуса [Рахилина 2009: 15–16], сменится интенсивной коллективной работой по созданию общерусского корпуса диалектной речи в составе НКРЯ.

Диалектный корпус, организованный как ***совокупность корпусов отдельных говоров***, обеспеченный ясной концепцией и инструктивными материалами, имеет реальную перспективу достижения репрезентативности в короткие сроки и может уже в ближайшее время стать научным источником, обладающим значительным эвристическим потенциалом.

References

1. *Grishina E. A.* 2009. Russian Multimedia Corpus: The Problems of Annotation [Multimediinyi Russkii Korpus (MURKO): Problemy Annotatsii]. Natsional'nyi Korpus Russkogo Iazyka: 2006-2008. Novye Rezul'taty I Perspektivy.
2. *Kriuchkova O. Iu.* 2007. Electronic Corpus of Russian Dialectal Speech, and the Principles of its Tagging [Elektronnyi Korpus Russkoi Dialektnoi Rechi I Printsipy ego Razmetki]. Izvestiia Saratovskogo Universiteta. Novaia Seriia. Filologiia. Zhurnalistika, 7 (1).
3. *Letuchii A. B.* 2005. Corpus of Dialectal Texts: Goals and Problems [Korpus Dialektnykh Tekstov: Zadachi I Problemy]. Natsional'nyi Korpus Russkogo Iazyka: 2003-2005. Rezul'taty I Perspektivy.
4. *Letuchii A. B.* 2009. Corpus of Dialectal Texts: Contents and Tagging Characteristics [Dialektnyi Korpus: Sostav I Osobennosti Razmetki]. Natsional'nyi Korpus Russkogo Iazyka: 2006-2008. Novye Rezul'taty I Perspektivy.
5. *Rakhilina E. V.* 2009. Corpus as a Creative Project [Korpus kak Tvorcheskii Proekt]. Natsional'nyi Korpus Russkogo Iazyka: 2006-2008. Novye Rezul'taty I Perspektivy.