

О.Ю. Крючкова, В.Е. Гольдин
O.Ju. Kryuchkova, V.E. Goldin

ПАРАМЕТРЫ ОБРАБОТКИ ТЕКСТОВ ДЛЯ РУССКОГО ДИАЛЕКТНОГО КОРПУСА

THE PARAMETERS OF TEXT PROCESSING FOR THE RUSSIAN DIALECT CORPUS

Аннотация. В докладе рассматриваются вопросы о представлении текстов в русском диалектном корпусе, о принципах расшифровки, разметки и метаразметки диалектного материала. Ставится актуальная для диалектного корпуса задача выхода за пределы узколингвистического представления материала, перспективах развития его лингво-культурологической направленности.

Ключевые слова. Корпус, диалект, текст, аннотирование.

Abstract. The report deals with the problems of representation of texts in the Russian dialect corpus, the principles of decoding, marking and meta-marking of dialect materials. The objective to overcome a narrowly linguistic representation of the material and the perspectives of developing its linguistic and cultural orientation are considered as very important ones.

Keywords. Corpus, dialect, text, annotation.

1. Введение

Опыт создания Саратовского диалектологического корпуса (СарДК) [Крючкова, Гольдин 2008, Крючкова, Гольдин 2011] и участие в наполнении диалектного подкорпуса Национального корпуса русского языка (ДК НКРЯ) служат основой выработанной нами общей концепции диалектного корпуса: 1) диалектный корпус носит не иллюстративный характер, а создается как полноценный научно-образовательный ресурс; 2) текстовая база корпуса строится как модель диалектной ком-

муникации; 3) она формируется с опорой на концепцию диалектного членения русского языка и представляет говоры как диалектные микросистемы; 4) диалектный корпус является корпусом лингво-культурологической направленности.

В настоящее время ни один из вопросов организации диалектного корпуса не получил общепринятого решения.

2. Принципы расшифровки и формы представления текстов в диалектном корпусе

Представлению диалектных текстов в корпусе предъявляются противоположные требования: обеспечить пользователям предельно легкую ориентацию в содержании текстов и максимально полно отразить не только лексические, грамматические, но и фонетические явления для научного анализа диалектной речи. Первой цели соответствует транскрипция, приближенная к орфографическому письму. Вторая цель требует максимального сохранения фонетической специфики говора.

В [Качинская 2011] предложен «новый стандарт подачи», предусматривающий 3 уровня представления текстов в корпусе: «текст в том виде, в каком он предоставлен изначально» («первоначальная транскрипция»), текст, «который представляет собой «облегченную» транскрипцию» и, наконец, получаемый автоматически «литературный подстрочник». Многоуровневая система представления диалектных текстов все же не решает, на наш взгляд, проблемы фонетической репрезентативности корпуса. Текст-источник рассматривается как основной источник фонетической информации в диалектном корпусе, но вряд ли символическая передача звучащей речи, выполненная разными исследователями, пригодна для фонетического анализа современного уровня.

Оптимальными формами корпусного представления диалектных текстов являются, как показывает наш опыт, звучащая речь (аудио- или видеозапись) и соотнесенная с ней вспо-

могательная транскрипция (полуорфографическая запись). Включение в корпус аудио- и видеозаписей диалектной коммуникации, соотнесенных с ее письменным представлением, в полном объеме сохраняет возможность фонетического анализа, а пользователю не навязываются фонетические квалификации автора расшифровки. В НКРЯ имеется успешный опыт создания устного корпуса со звуковой составляющей, который важно было бы использовать и при разработке диалектного корпуса.

Принятие предлагаемых нами форм корпусного представления диалектных текстов требует тщательной разработки принципов вспомогательной транскрипции и определения форм соотнесения аудио- и видеозаписей с символьными расшифровками текстов. Разработчиками СарДК создана основа такого стандарта. Его наличие и регулярное коллективное обсуждение трудных случаев расшифровки обеспечивают «на выходе» единство формата письменного представления в корпусе звучащей диалектной речи. Применяемая в СарДК вспомогательная транскрипция ограничивает передачу фонологически незначимых явлений, различного рода позиционных и комбинаторных особенностей реализации фонем. Информацию о них пользователь получает из первоисточника – звуковой фиксации диалектной речи.

3. Аннотирование текстовой базы корпуса

3.1. Морфологическая разметка

При аннотировании диалектных текстов в ДК НКРЯ использована сложная система дифференциальных помет, описывающая типы «отклонений» от литературной нормы (*dialflex* – «диалектное окончание», *dialmorph* – «диалектная морфология», *dialstem* – «диалектная основа», *gend* – «диалектные особенности в сфере рода»). Стратегия детальной характеристики словоформ, не отвечающих литературному

стандарту, ведет к постоянным уточнениям и изменениям в системе помет, но не решает проблемы функциональности разметки. Необходимо, на наш взгляд, отказаться в разметке от конкретной квалификации диалектных особенностей, ограничившись общей пометой, указывающей на отличие текстовой формы от литературного стандарта. Соотнесение диалектных (и иных нелитературных) форм с литературными необходимо в корпусе, но имеет значение не для квалификации таких форм, а для оптимальной организации поиска как по нелитературным формам, так и по их литературным соответствиям. В отличие от ДК НКРЯ в СарДК принят более простой и функционально оправданный способ установления диалектно-литературных соответствий: в грамматическую разметку введена зона литературного соответствия, идущая вслед за зоной начальной формы слова (ср: *one{one(они)=S-PRO,мн,од=им=nstand}*).

Особого внимания заслуживает вопрос о целесообразности специальной маркировки и способах разметки различных видов неоднословных и идиоматических единиц. Для диалектного корпуса такая маркировка необходима, поскольку набор соответствующих единиц заранее неизвестен пользователю и получить их систематическую выборку путем стандартных запросов, базирующихся на пословной разметке, невозможно, ср.: *мало чё (мало ли что), звон дают (звонят), ноги отсеклись (ноги отнялись), пошли все по нём (пошли все в него), быть на этом (так тому и быть), на причине (быть стельною – о корове), после времени (задним числом)*. В НКРЯ и в его диалектном подкорпусе специальная разметка неоднословных единиц не проводится, хотя их перечень, выделенный из Основного подкорпуса НКРЯ, представлен в отдельном словнике (Словарь неоднословных лексических единиц (оборотов)). В СарДК разметка неоднословных единиц ведется с учётом специфики разных типов таких единиц: раздельно-оформленные структурные единицы (союзы, частицы, наречия,

вводные слова, междометия типа *a to, как же, все равно, в общем, батюшки мое*), раздельнооформленные словоформы (типа *кое у кого*), идиоматические выражения.

3.2. Метаразметка диалектных текстов

Для корпусов устной речи, в которой темы и жанры коммуникации редко обозначаются самими коммуникантами, значимыми типами разметки являются тематическая и жанровая. Их особенность в диалектном корпусе, как вообще в корпусах устной разговорной речи, состоит в характеристике в н у т р е н н е г о членения речевых фрагментов (текстов), поскольку устная речь принципиально политематична и полижанрова.

В ДК НКРЯ эта специфика не учитывается, и тематическая разметка ведется, как в других корпусах, по корпусу в целом. В СарДК, где единицей хранения является значительный по объему политематичный и полижанровый речевой фрагмент (максимальное приближение к формату естественной коммуникации – один из важных принципов построения СарДК), проводится тематическая и жанровая разметка каждого диалектного текста. В перспективе она даст возможность исследовать жанрово-тематическую структуру диалектной коммуникации, выявлять соотношение различных тем и жанров в ее составе.

Диалектный корпус в силу специфики традиционной культуры не может быть узко лингвистическим, его необходимо создавать как лингво-культурологический корпус. Принципиальное значение для корпусов лингвокультурологической направленности имеет структура метаразметки.

Используемых ДК НКРЯ в настоящее время параметров метаразметки явно недостаточно для понимания и анализа диалектной коммуникации. 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.

Важными оказываются сохраняемые в корпусе 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 метаразмтку текстов в СарДК.

Проектом «Новый стандарт подачи» [Качинская 2011: 243] предусматривается значительное изменение структуры метаразмтки диалектных текстов, включение в метаописание достаточно подробных сведений об информанте (местный, приезжий, откуда приехал, как давно живет в данном населенном пункте, род занятий, конфессия и др.). Важным представляется также дополнение метаразмтки фонетическим комментарием к тексту, характер которого требует специального обсуждения.

4. Заключение

Репрезентативность диалектного корпуса во многом определяется форматом представления в нём диалектной речи и характером аннотирования текстов, однако существенными являются также вопросы общего строения диалектного корпуса, критерии отбора и структурирования речевого материала. Важное значение для диалектологического корпуса имеет также решение вопроса о включении в его состав блоков нелингвистической информации, поскольку диалектный корпус в наибольшей мере тяготеет к корпусам лингвокультурологической направленности.

Примеры удачного выхода за пределы узко лингвистического представления материала (выхода в соответствии со

спецификой материала) есть в НКРЯ (ср., напр., поэтический корпус, мультимедийный корпус). По-видимому, это означает, что есть смысл обсуждать общие принципы и возможности культурологических и иных «приращений» к речевой составляющей корпусов в составе НКРЯ. Для диалектного корпуса этот вопрос имеет особую остроту. Разработка стандартов включения в диалектологический корпус экстралингвистической информации (этнографической, этнокультурной, сопровождение текстов фотоматериалами) действительно было бы важным шагом в развитии корпусов этого типа.

Литература

1. *Качинская И.Б.* (2011), Диалектный подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место. Русская устная речь. Саратов, с. 240–249.

2. *Крючкова О.Ю., Гольдин В.Е.* (2008), Текстовый диалектологический корпус как модель традиционной сельской коммуникации. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008», с. 268–273.

3. *Крючкова О.Ю., Гольдин В.Е.* (2011), Корпус русской диалектной речи: концепция и параметры оценки. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2011», с. 359–367.

References

1. *Kachinskaja I.B.* (2011), Dialektnyj podkorpus NKRJA. Novyj standart podachi. Novoje rabocheje mesto. [The Dialect Subcorpus of The National Corpus of The Russian Language. A new standard of representation. A new working place.]. Russkaja ustnaja rech. [Russian spoken language]. Saratov, pp. 240–249.

2. *Kryuchkova O.Ju., Goldin V.E.* (2008), Tekstovyj dialektologicheskij korpus kak model' tradicionnoj sel'skoj kommu-

nikacii [Textual dialect corpus as a model of traditional rural communication]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2008» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2008»], pp. 268–273.

3. Kryuchkova O.Ju., Goldin V.E. (2011), Korpus ruskoj dialektnoj rechi: koncepcija i parametry ocenki [Corpus of Russian dialect speech: concept and parameters of evaluation]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2011» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2011»], pp. 359–367.

Крючкова Ольга Юрьевна

Саратовский государственный университет им. Н.Г. Чернышевского (Россия)

Kryuchkova Olga

Saratov State University (Russia)

E-mail: vpks@rambler.ru

Гольдин Валентин Евсеевич

Саратовский государственный университет им. Н.Г. Чернышевского (Россия)

Goldin Valentin

Saratov State University (Russia)

E-mail: goldinve@yandex.ru