***Некрасова Г. А.**, кандидат филологических наук, старший научный сотрудник ИЯЛИ КНЦ УрО РАН.*

Мультимедийная база данных коми диалектов как ресурс сохранения коми культуры

Коми язык, представляющий собой единство трех территориально-языковых разновидностей: коми-зырянской, коми-пермяцкой и коми-язьвинской, распространен на территории нескольких административных образований Российской Федерации: Республика Коми, Тюменская, Мурманская, Архангельская, Кировская области, Пермский край. Каждый говор, диалект, каждое наречие – это важнейшая часть коми языка, это его культурно-территориальные разновидности. Поэтому изучение особенностей диалектов, истории их формирования и развития является составной частью изучения истории коми народа.

Исследование коми диалектов началось со второй половины XIX в., планомерный и систематический характер оно получило в конце 40-х годов прошлого столетия. Большой вклад в изучение коми диалектов внеси как зарубежные (М. Кастрен, Ю. Вихманн, Д. Фокош-Фукс, Т. Уотила), так и российские (В. И. Лыткин, А. С. Сидоров, М. А. Сахарова, Н. А. Колегова, В. А. Сорвачева, Т. И. Жилина) исследователи (подробно об истории изучения коми диалектов в [16, *29-39*]). За годы исследования коми языка был собран обширный и разнообразный материал почти по всем коми диалектам, на основе которого были подготовлены текстовые и лексикографические публикации, в числе которых уникальные издания «Образцы коми-зырянской речи» [15] и «Сравнительный словарь коми-зырянских диалектов» [22], текстовые материалы и словари Д. Фокоша-Фукса [24-25], Т. Уотилы [26-29]. В коми диалектологии описаны все зырянские, большинство пермяцких диалектов и

коми-язьвинское наречие [3-10; 12; 14; 17-21], имеется ряд обобщающих работ [1-2; 13; 16]. В настоящее время продолжается фиксация материала по всем территориально-языковым разновидностям коми языка. Почти ежегодно организуются экспедиции студентами Сыктывкарского государственного университета (СыктГУ) и сотрудниками Института языка, литературы и истории Коми научного центра Уральского отделения (КНЦ УрО) РАН. Однако обработка диалектного материала в большинстве случаев ограничивается расшифровкой звуковых источников и транскрибированием текстов. Накопленный уникальный языковой материал, насчитывающий несколько сотен аудиокассет, томов рукописного материала, несколько сотен тысяч карточек, хранится в фондах научного архива КНЦ УрО РАН, сектора языка ИЯЛИ, а также в фондах кафедры коми и финно-угорской филологии филологического факультета СыктГУ. Собранный диалектный материал мало доступен широкому кругу исследователей, а также не может быть использован для всех типов лингвистических исследований, особенно для исследований фонетики и акцентуации диалектов.

Особую значимость для диалектологии имеет создание диалектных текстовых корпусов, поэтому одним из главных проектов по сохранению и развитию коми языка должна была стать разработка проекта по созданию Электронной базы коми диалектов (ЭБКД), включающей два основных вида электронных ресурсов – корпус текстов и словарь.

Основу электронной базы коми диалектов должен составить программно обеспеченный электронный текстовый корпус, являющийся наиболее надежной формой хранения диалектных текстов. Лингвистическим корпусом называют собрание текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой. К настоящему времени лингвистика уже обогатилась электронными языковыми ресурсами, к числу крупных корпусов относятся Британский национальный корпус, Чешский национальный корпус, Национальный корпус русского языка. Ведется создание электронных ресурсов отдельных финно-угорских языков. Так, Corpus of Estonian Written Texts содержит эстонские письменные тексты, опубликованные с 1983 по 1987 гг.: газетные тексты, художественная литература, научные и научно-популярные тексты и др. Тексты снабжены метаразметкой. Поиск по корпусу не предусмотрен, для просмотра можно скачать или содержащийся в одном файле корпус без разметки, или разбитый на отдельные файлы по типам текстов размеченный корпус.

При создании корпуса коми текстов необходимо учитывать международный опыт создания электронных баз конкретных язы-

ков. В рамках проекта по документации коми диалектов могут быть использованы разработанный исследовательской группой А. Е. Кибрика стандарт представления языкового материала и программная среда для создания и использования мультимедийных языковых ресурсов для малых языков. «Разработанный на современном уровне корпус текстов для малого языка должен отвечать следующим требованиям:

· корпус может публиковаться на бумаге, но обязательно должен существовать в виде электронной базы данных, допускающей обновление и пополнение;

· тексты должны быть представлены не только в транскрипции, уже представляющей результат определенной интерпретации исследователем языковых данных, но и в исходном виде, т. е. в (видео и) аудиозаписях;

· тексты должны нести максимум лингвистической и иной разметки (аннотации), в том числе обязательно поморфемное глоссирование, а также, в зависимости от ресурсов, которыми располагает исследовательская группа, дополнительные слои морфонологической, просодической, синтаксической, частеречной, семантической разметки; текстовые, метатекстовые и энциклопедические комментарии;

· аннотация должна быть в высокой степени стандартизована для того, чтобы облегчить поиск сходных явлений в разных языках, описанных разными исследователями;

· пользователь должен иметь возможность выбрать для отображения только интересующие его компоненты (слои) информации;

· корпус должен поддерживать обработку самых разнообразных поисковых запросов пользователя, включая поиск по различным слоям разметки (например – и в первую очередь – по грамматическим глоссам и пр.);

· корпус должен являться открытым Интернет-ресурсом, чтобы любой заинтересованный пользователь мог легко к нему обратиться» [11, *232-233*].

Корпус коми диалектов должен представлять собой коллекцию электронных текстов, снабженных лингвистической и метатекстовой информацией. Основная задача диалектного корпуса заключается в представлении каждого диалекта как территориальной разновидности коми языка. Поэтому каждый отдельный диалект, а также говор в составе диалекта должен образовать самостоятельный подкорпус в составе корпуса. Основным принципом формирования текстовой базы корпуса должен стать принцип полного и адекватного отражения в корпусе специфики диалекта, что предполагает наполнение каждого подкорпуса разнообразным значительным по объему текстовым материалом, репрезентирующим

различные формы речи (диалог, монолог); важнейшие типы речи (бытовую, фольклорную, официальную); социальную дифференциацию носителей говора (по полу, возрасту, уровню образования). Текст необходимо представить в двух видах: в виде звукового модуля – аудиофайла, в виде графического модуля – графического изображения транскрипции, снабженной переводом на русский язык.

Корпус должен содержать дополнительную информацию о свойствах входящих в него текстов (разметку, или аннотацию). Это главная характеристика корпуса, отличающая корпус от простых коллекций текстов. Каждый текст должен иметь лингвистическую и экстралингвистическую разметку. На первоначальном этапе обработки текста достаточной является минимальная метатекстовая разметка, характеризующая текст в целом. Параметрами метаразметки должны стать нелингвистические сведения о тексте и о диалекте. В информацию о тексте необходимо включить сведения об информантах, о времени, месте записи (краткая история населенного пункта, описание специфических природных особенностей, микротопонимы), о конкретной ситуации общения, об адресатах речи, упоминаемых лицах и их отношении к информанту, о времени событий в повествовании. Информация о диалекте (говоре) должна содержать сведения о составе диалекта, истории формирования диалекта, краткая история изучения диалекта, библиографию об изучении диалекта. Вербальный ряд информации необходимо дополнить графической информацией, включив в корпус карты, схемы, фотографии. Часть данной информации может быть соотнесена в корпусе с текстовыми модулями, другая часть образовать отдельный информационный блок. Поисковые запросы могут осуществляться в соответствии с каждым из параметров метаразметки. Формат представления информации в корпусе необходимо разработать с учетом существующих стандартов для кодирования корпусов (TEI, XCES, EAGLES). На последующих этапах корпус должен быть дополнен системой лингвистической (морфологической, акцентной, синтаксической и семантической) разметки. Большинство специалистов отмечают необходимость перевода текстов на язык-посредник, а также поморфемного глоссирования текстов. Однако, как отмечают Е. А. Кибрик и др., «единого унифицированного формата глоссирования и, в целом, представления текстов сейчас не существует. Различия касаются не только инвентаря грамматических глосс, но также и количества и состава необходимых слоев репрезентации. Это связано как с объективными научно содержательными проблемами (неизоморфность грамматической структуры различных языков, различия в степени прозрачности морфонологических процессов и т. п.), так и с организационными

(отсутствие единого координирующего центра или стандарта)» [11, *233*]. В качестве основного справочника при глоссировании текстов необходимо использовать научную грамматику коми языка [23], где наиболее полно описаны грамматические категории. Следуя рекомендациям правил глоссирования в отношении используемых символов-разделителей, а также сокращений, принятых для обозначения тех или иных грамматических категорий, при описании языковых единиц необходимо внести дополнения (глоссы) для не включенных в стандарт категорий, например, категории степеней действия глагола, приблизительно-местных падежей. В итоге языковые данные должны быть обработаны и проиндексированы таким образом, чтобы ими можно было пользоваться и в последующие столетия, а документация должна быть архивирована так, чтобы ее легко можно было сохранить и при необходимости перенести на новые носители информации.

Электронная форма представления диалектных текстов повышает сохранность собранного уникального материала, создает возможность для более свободного доступа лингвистов различной специализации к диалектному материалу, позволяющему наблюдать реальные отношения между языковыми единицами в потоке диалектной речи. Программное обеспечение корпуса позволит каждому исследователю при минимальных затратах усилий самостоятельно создавать на основе корпуса полные базы данных в соответствии со своими исследовательскими задачами, классифицировать материал на основании отдельных параметров и их комплексов. Электронные ресурсы должны стать источником исследовательских проектов, диссертационных и дипломных работ, программно-методических комплексов в общеобразовательной и вузовской системах преподавания. Они дадут более широкие возможности для проведения сравнительно-исторических, сопоставительных и типологических исследований.

Важной составляющей ЭБКД должен стать электронный словарь коми диалектов, который необходимо сформировать путем внесения в него всех лексем, содержащихся в источниках, составляющих электронный корпус коми диалектов. В основу словаря могут лечь [22], а также «Словарь диалектов коми языка», составляемый сотрудниками сектора языка ИЯЛИ КНЦ УрО РАН. Отличительной чертой словаря должно стать наличие иллюстративных контекстов, направленных на раскрытие семантической структуры слова и описание всех его значений. Каждое значение должно быть проиллюстрировано примерами из текстов, составляющих электронный корпус коми диалектов. Все лексемы в словаре должны быть паспортизированы. По составленному эталонному словнику, организованному по тезаурусному принципу и содержащему в

основном базовую лексику коми диалектов, можно сделать запись материалов для звукового словаря. Для историко-типологических исследований важным является включение в состав словника одного из списков Сводеша.

Для осуществления проекта ЭБКД необходимо продолжить фиксацию речи носителей диалектов разных поколений, в разной степени владеющих языком. Это даст возможность проследить динамику изменения языковой структуры в ситуации коми-русского билингвизма. В первую очередь необходимо продолжить документацию говоров и диалектов, которые находятся на грани исчезновения. К числу таких территориальных разновидностей следует отнести коми-язьвинское наречие и те говоры, которые составляют пограничные зоны с северно-русскими говорами. Для фиксации диалектного материала необходимо привлечь как можно большее количество участников документации. Традиционно сбор диалектного материала проводится, как правило, единичными лингвистами или же группой студентов, которые производят и расшифровку материала. Представляется возможным, включить в эту работу носителей диалекта, предварительно обучив их соответствующим методам фиксации материала и снабдив необходимыми техническими средствами. Только создание электронной базы коми диалектов позволит реставрировать и сохранить уникальные аудио и рукописные материалы, накопленные коми языковедами в течение последнего столетия.

Литература

1. *Баталова, Р. М.* Коми-пермяцкая диалектология / Р. М. Баталова. – М.: Наука, 1975. – 252 с.

2. *Баталова, Р. М.* Ареальные исследования по восточным финно-угорским языкам (коми языки) / Р. М. Баталова. – М.: Наука, 1982. – 167 с.

3. *Баталова, Р. М.* Оньковский диалект коми-пермяцкого языка. Унифицированное описание диалектов уральских языков / Р. М. Баталова. – М., 1990. – 205 с.

4. *Баталова, Р. М.* Нижнеиньвеньский диалект коми-пермяцкого языка / Р. М. Баталова. – М.-Гамбург, 1995. – 197 с.

5. *Баталова, Р. М.* Кудымкарско-иньвеньский диалект коми-пермяцкого языка / Р. М. Баталова. – М.-Гамбург, 2002. – 168 с.

6. *Дмитриева, Р. П.* Косинско-камский диалект коми-пермяцкого языка (фонетика, морфология): Дис... канд. филол. наук / Р. П. Дмитриева. – Йошкар-Ола, 1998. – 195 с.

7. *Жилина, Т. И.* Верхнесысольский диалект коми языка / Т. И. Жилина. – М.: Наука, 1975. – 268 с.

8. *Жилина, Т. И.* Вымский диалект коми языка / Т. И. Жилина. – Сыктывкар: Пролог, 1998. – 439 с.

9. *Жилина, Т. И.* Лузско-летский диалект коми языка / Т. И. Жилина. – М.: Наука, 1985. – 272 с.

10. *Жилина, Т. И., Бараксанов, Г. Г.* Присыктывкарский диалект и коми литературный язык / Т. И. Жилина, Г. Г. Бараксанов. – М.: Наука, 1971. – 276 с.

11. *Кибрик, А. Е., Архипов, А. В., Даниэль, М. А., Кодзасов, С. В., Майерс, Том, Нахимовский, А. Д.* Технологии обработки языковых данных в документировании малых языков // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.) / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. – М.: Изд-во РГГУ, 2007. – С. 231-235.

12. *Колегова, Н. А., Бараксанов, Г. Г.* Среднесысольский диалект коми языка / Н. А. Колегова, Г. Г. Бараксанов. – М.: Наука, 1980. – 226 с.

13. *Лыткин, В. И.* Диалектологическая хрестоматия по пермским языкам (с обзором диалектов и диалектологическим словарем) / В. И. Лыткин. – М. Изд-во АН СССР, 1955. – 128 с.

14. Лыткин, В.И. Коми-язьвинский диалект / В. И. Лыткин. – М.: Изд-во АН СССР, 1961. – 228 с.

15. Образцы коми-зырянской речи. – Сыктывкар: Коми кн. изд-во, 1971. – 311 с.

16. *Сажина, С. А.* Сравнительная морфология коми-зырянских диалектов (именные части речи). Ареальный аспект исследования. Дис. ... канд. филол. наук / С. А. Сажина. – Сыктывкар, 2004. – 282 с.

17. *Сахарова, М. А., Сельков, Н. Н.* Ижемский диалект коми языка / М. А. Сахарова, Н. Н. Сельков. – Сыктывкар: Коми кн. изд-во, 1976. – 288 с.

18. *Сахарова, М. А., Сельков, Н. Н., Колегова, Н. А.* Печорский диалект коми языка / М. А. Сахарова, Н. Н. Сельков, Н. А. Колегова. – Сыктывкар: Коми кн. изд-во, 1976. 153 с.

19. *Сорвачева, В. А.* Нижневычегодский диалект коми языка/ В. А. Сорвачева. – М.: Наука, 1978. 227 с.

20. *Сорвачева, В. А., Безносикова, Л. М.* Удорский диалект коми языка / В. А. Сорвачева, Л. М. Безносикова. – М.: Наука, 1990. – 283 с.

21. *Сорвачева, В. А., Сахарова М. А., Гуляев, Е.С.* Верхневычегодский диалект коми языка / В. А. Сорвачева, М. А. Сахарова, Е. С. Гуляев. – Сыктывкар: Коми кн. изд-во, 1966 (Историко-филологический сборник. Вып. 10). – 256 с.

22. Сравнительный словарь коми-зырянских диалектов / Под ред. В. А. Сорвачевой. – Сыктывкар: Коми кн. изд-во, 1961. – 90 с.

23. Ӧнія коми кыв. Морфология = Современный коми язык. Морфология / В. М. Лудыкова, Г. А. Некрасова, Э. Н. Попова, Г. В. Федюнева, Е. А. Цыпанов. – Сыктывкар: Коми кн. изд-во, 2000. – 544 с.

24. *Fokos-Fuchs, D. R.* Volksdichtung der Komi (Syrjanen) / D. R. Fokos-Fuchs. – Budapest, 1951. – 472 S.

25. *Fokos-Fuchs, D. R.* Syrjanisches Worterbuch. I, II / D. R. Fokos-Fuchs. – Budapest: Akademiai Kiado, 1959. – 1654 S.

26. *Uotila, T. E.* Syrjanische Texte. Band I. Komi-permjakisch / T. E. Uotila. – Helsinki. 1985. – 297 S.

27. *Uotila, T. E.* Syrjanische Texte. Band II / T. E. Uotila. – Helsinki, 1986. – 242 S.

28. *Uotila, T. E.* Syrjanische Texte. Band III / T. E. Uotila. – Helsinki, 1989. – 402 S.

29. *Uotila, T. E.* Syrjanische Texte. Band IV / T. E. Uotila. – Helsinki, 1995. – 297 S.