Электронный диалектный корпус как ресурс сохранения и изучения коми диалектов

Г. А. Некрасова,
кандидат филологических наук, старший научный сотрудник сектора языка Института языка, литературы и истории Коми научного центра Уральского отделения РАН (г. Сыктывкар, РФ)

Коми язык, представляющий собой единство трех территориально-языковых разновидностей: коми-зырянской, коми-пермяцкой и коми-язьвинской, – распространен на территории нескольких административных образований Российской Федерации: Республики Коми, Тюменской, Мурманской, Архангельской, Кировской областей, Пермского края. Каждый говор, диалект, каждое наречие – важнейшая часть коми языка. Поэтому изучение особенностей диалектов, истории их формирования и развития является составляющей изучения истории коми народа.

Исследование коми диалектов началось во второй половине XIX в., планомерный и систематический характер оно получило в конце 40-х гг. прошлого столетия. Большой вклад в это дело внесли 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) ученые (подробнее см.: [16, *29–39*]). За прошедшие годы был собран обширный и разнообразный материал почти по всем коми диалектам, на основе которого были подготовлены текстовые и лексикографические публикации. В их числе – 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 студентов Сыктывкарского государственного университета и сотрудников 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 и томов рукописного материала, несколько сотен тысяч карточек, хранится в фондах научного архива КНЦ УрО РАН, сектора языка Института языка, литературы и истории, а также в фондах кафедры коми и финно-угорской филологии филологического факультета университета и мало доступен широкому кругу исследователей, и в первую очередь тем, кто занимается фонетикой и акцентуацией диалектов.

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

Основу ЭБКД должен составить программно обеспеченный электронный текстовый корпус, являющийся наиболее надежной формой хранения диалектных текстов. В лингвистике корпусом принято называть собрание текстов, собранных в соответ-ствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной

© Некрасова Г. А., 2010

поисковой системой. К настоящему времени лингвистика уже обогатилась электронными языковыми ресурсами, к числу крупнейших из которых 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 г.: газетные тексты, художественную литературу, 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.

Исследование коми диалектов началось во второй половине XIX в., планомерный и систематический характер оно получило в конце 40-х гг. прошлого столетия.

При создании корпуса коми текстов необходимо учитывать международный опыт подобной деятельности. В рамках проекта могут быть использованы разработанный исследовательской группой А. Е. Кибрика стандарт представления языкового материала и программная среда для создания и применения мультимедийных языковых ресурсов для малых языков. Корпус 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;

В коми диалектологии описаны все зырянские, большинство пермяцких диалектов и коми-язьвинское наречие, имеется ряд обобщающих работ.

– 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 – это коллекция электронных текстов, снабженных лингвистической и метатекстовой информацией. Его основная задача заключается в представлении каждого диалекта как территориальной разновидности коми языка. Каждый отдельный диалект, а также говор в составе диалекта должен образовывать самостоятельный подкорпус в составе корпуса. Главными при формировании текстовой базы корпуса должен стать принцип полного и адекватного отражения 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.

Главная характеристика корпуса, отличающая его от простых коллекций текстов, заключается в наличии дополнительной информации о свойствах входящих в него текстов (разметки, или аннотации). 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 его составе, истории формирования, краткой истории 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 – образовывать отдельный информационный блок. Поисковые запросы осуществляются в соответствии с каждым из параметров метаразметки. Формат представления информации в корпусе разрабатывается с учетом существующих стандартов для кодирования корпусов (TEI, XCES, EAGLES).

Основу электронной базы коми диалектов должен составить программно обеспеченный электронный текстовый корпус, являющийся наиболее надежной формой хранения диалектных текстов.

На последующих этапах корпус дополняется системой лингвистической (морфологической, акцентной, синтаксической и семантической) разметки. Большинство специалистов отмечают необходимость перевода текстов на язык-посредник, а также поморфемного глоссирования текстов. Однако, по мнению А. Е. Кибрика 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 целесообразно 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 будущем, 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, обеспечивает условия для лингвистов различной специализации более свободного доступа к диалектному материалу, позволяет наблюдать реальные отношения между языковыми единицами в потоке диалектной речи, при минимальных затратах усилий самостоятельно создавать 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 создадут более широкие возможности для проведения сравнительно-исторических, сопоставительных и типологических исследований.

Важной составляющей ЭБКД должен стать электронный словарь коми диалектов, который формируется путем внесения в него всех лексем, содержащихся в источниках, составляющих электронный корпус коми диалектов. В основу словаря могут лечь материалы Сравнительного словаря коми-зырянских диалектов [22], а также Словаря диалектов коми языка, составляемого сотрудниками сектора языка Института языка, литературы и истории КНЦ УрО РАН. Отличительной чертой электронного 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 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 языковой структуры в ситуации коми-русского

Главная характеристика корпуса, отличающая его от простых коллекций текстов, заключается в наличии дополнительной информации о свойствах входящих в него текстов.

Создание электронной базы коми диалектов позволит реставрировать и сохранить уникальные аудио- и рукописные материалы, накопленные коми языковедами в течение последнего столетия.

билингвизма. В первую очередь внимание к себе требуют говоры и диалекты, находящиеся на грани исчезновения. К числу таких территориальных разновидностей относятся коми-язьвинское наречие и те говоры, которые составляют пограничные зоны с северно-русскими говорами. Для фиксации диалектного материала надо привлечь как можно большее количество участников документации. Традиционно сбор диалектного материала проводится 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.

КЛЮЧЕВЫЕ СЛОВА

электронная база диалектов; электронные словари; корпус текстов; коми язык

dialects electronic database; electronic dictionaries; texts corpus; the Komi language

KEYWORDS

Поступила 14.12.2009

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

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.* Присыктывкарский диалект и коми литературный язык / Т. И. Жилина, Г. Г. Бараксанов. – М. : Наука, 1971. – 276 с.
11. *Кибрик, А. Е.* Технологии обработки языковых данных в документировании малых языков / А. Е. Кибрик, А. В. Архипов, М. А. Даниэль и др. // Компьютерная лингвистика и интеллектуальные технологии : тр. междунар. конф. «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.) / под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. – М., 2007. – С. 231–235.
12. *Колегова, Н. А.* 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.* Ижемский диалект коми языка / М. А. Сахарова, Н. Н. Сельков. – Сыктывкар : Коми кн. изд-во, 1976. – 288 с.
18. *Сахарова, М. 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.* Удорский диалект коми языка / В. А. Сорвачева, Л. М. Безносикова. – М. : Наука, 1990. – 283 с.
21. *Сорвачева, В. А.* Верхневычегодский диалект коми языка / В. А. Сорвачева, М. А. Сахарова, Е. С. Гуляев. – Сыктывкар : Коми кн. изд-во, 1966. – 256 с. (Ист.-филол. сб. ; вып.10).
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 (Syrjänen) / D. R. Fokos-Fuchs. – Budapest, 1951. – 472 S.
25. *Fokos-Fuchs, D. R.* Syrjänisches Wörterbuch. I, II / D. R. Fokos-Fuchs. – Budapest : Akadémiai Kiadó, 1959. – 1654 S.
26. *Uotila, T. E.* Syrjänische Texte. Bd. 1. Komi-permjakisch / T. E. Uotila. – Helsinki, 1985. – 297 S.
27. *Uotila, T. E.* Syrjänische Texte. Bd. 2 / T. E. Uotila. – Helsinki, 1986. – 242 S.
28. *Uotila, T. E.* Syrjänische Texte. Bd. 3 / T. E. Uotila. – Helsinki, 1989. – 402 S.
29. *Uotila, T. E.* Syrjänische Texte. Bd. 4 / T. E. Uotila. – Helsinki, 1995. – 297 S.