Автоматическая обработка диалектного материала для включения в электронный текстовый корпус

Батраева, И. А.1, Трещева, Е. Г.2

1 batraevaia@info.sgu.ru, 2 treshyova@gmail.com1,2Саратовский государственный университет им. Н.Г. Чернышевского, Саратов, Россия

В работе рассматриваются принципы и основные этапы подготовки диалектного текста для включения в базу мультимедийного текстового корпуса СарДК, разрабатываемого на базе кафедры теории, истории языка и прикладной лингвистики СГУ.

Ключевые слова: корпусная лингвистика, корпус диалектных текстов, разметка

В настоящее время лингвистика как наука развивается в рамках новой парадигмы – говорят даже о том, что наука переживает «корпусную революцию»: исследователи все чаще опираются не на собранные вручную тексты, а на данные статистически значимых и репрезентативных текстовых массивов – электронных текстовых корпусов. Такие собрания текстов, сбалансированных по составу и снабженных дополнительной информацией (разметкой), позволяют представить язык (или его вариант) в его реальном употреблении и служат, таким образом, достоверной моделью некоторой коммуникативной сферы.

На базе кафедры теории, истории языка и прикладной лингвистики разрабатывается проект «Саратовский диалектный корпус» (далее СарДК) [1]. В задачи этого корпуса входит моделирование диалектной коммуникации в пределах четырех русских говоров (говоры сел Белогорное, Земляные Хутора и Орлов Гай Саратовской области и куста сел Мегра Вологодской области). Корпус представляет собой мультимедийный ресурс и включает, помимо текстовой информации, звуковые и видеофайлы. На настоящий момент собран большой массив диалектных материалов, разработаны структура корпуса и алгоритм поиска по размеченному тексту [2]. В задачи участников проекта сейчас входит создание веб-интерфейса СарДК, взаимодействующего с поисковым механизмом, ориентированным на лингвистические особенности диалектного материала.

При создании электронного текстового корпуса важен вопрос о его структурной единице. В СарДК этой единицей является модуль, соответствующий одной диалектной аудио- или видеозаписи. Это решение является теоретически обоснованным в рамках концепции коммуникативной диалектологии, так как «сохранение целого текста, представляющего естественную речь носителя народной речевой культуры, дает больше возможностей для изучения особенностей диалектной коммуникации, ее когнитивно-дискурсивной специфики» [3: 73].

Подготовка модуля СарДК предполагает следующие этапы, опирающиеся как на автоматический, так и ручной способ обработки диалектного материала:

1. Создание звукового или видеофайла, фиксирующего беседу с носителем диалекта (запись диалектных текстов / оцифровка аналоговых носителей);

2. Подготовка метаинформации: сведения об информанте (ФИО, дата рождения, пол, род занятий), а также информация о записи текста (дата записи; место и условия коммуникации; тематические и жанровые характеристики беседы; данные о собирателях, об ответственных за подготовку модуля); биографические данные информанта; фотографии и карты;

3. Расшифровка аудио- или видеозаписи (беседа с диалектоносителем переводится в текстовое представление);

4. Тематическая и жанровая разметка (ее наличие является важной особенностью корпуса, см. [3]);

5. Нарезка аудио- или видеофайла на фрагменты (не превышающие по длительности 30 с) для поисковой выдачи, расстановка маркеров границ этих фрагментов в соответствующем тексте расшифровки;

6. Автоматическая морфологическая разметка, в ходе которой текстоформам приписываются грамматические признаки;

7. Ручное снятие морфологической неоднозначности, добавление в разметку дополнительных сведений о текстоформе, обусловленных спецификой диалектной коммуникации (литературное соответствие, сведения о нестандартном употреблении, о вхождении в состав неоднословной единицы – идиоматического выражения или аналитической формы). Результат первых семи этапов подготовки текстов представлен на рис. 1.

8. Подготовка текстового представления в XML-ориентированном формате для отображения размеченного текста в веб-интерфейсе, нарезка файла с размеченным текстом на фрагменты, соответствующие звуковым / видеофрагментам, подготовленным на более раннем этапе.

Таким образом, готовый модуль содержит в себе заключенную в отдельные файлы разностороннюю информацию об основной единице корпуса – диалектной записи.

```
&1 #11@1 %А вот раньше было, там не разрешали крест носить и вообще
веровать – как тогда было?% &2 а{а(а)=CONJ}
всё{весь(весь)=A,idiom=ед,сред,им}+равно{равно(равно)=PRAEDIC,idiom}//
всё{весь(весь)=A,idiom=ед,сред,им}+равно{равно(равно)=PRAEDIC,idiom}
пересилили{пересилить(преодолеть)=V=сов,изъяв,прош,мн=nstand}/
дочка{дочка(дочка)=S,жен,од=ед,им}/…
```

Рис. 1. Пример диалектного текста с морфологической ({}), жанрово-тематической (#@) разметкой, а также с маркерами слов диалектолога (%) и границ аудио (&)

При выполнении задачи обращения к размеченному диалектному тексту через поисковый механизм удобно, чтобы информация о тексте любого уровня (в СарДК это лексико-грамматическая, жанрово-тематическая информация, метаинформация о тексте и носителях диалекта) была представлена в одном файле, но в то же время отвечала требованиям простоты и структурированности. Коллективом разработчиков СарДК было принято решение опираться на XML-ориентированный формат разметки, принятый в Национальном корпусе русского языка (НКРЯ) [4].

Этот формат удобен, в том числе тем, что позволяет учесть структурно-лингвистические особенности текста. Так, в разметке сохраняется иерархия текстовых элементов: уровень текстового фрагмента, характеризующегося одними и теми же метахарактеристиками (говорящий, тема и жанр), иерархически подчиненный ему уровень отдельных текстовых единиц (слово, неоднословная структурная единица, знак пунктуации). Кроме того, гибкость формата позволяет вводить новые теги и их параметры. К примеру, к стандартным грамматическим признакам, размечаемым по стандарту НКРЯ (см., например, [5]), добавлены пометы *idiom*, *analit*, *nstand*, *ind*, маркирующие соответственно такие особенности текстоформ как вхождение в состав идиоматического сочетания (*всё равно*, *на всякий случай*) или аналитической формы (*буду делать*), нестандартность леммы / грамматической формы (*гаманок* («кошелек»), *на рек**им***), окказионализм (*грунт* вместо *гурт*) и др. Важной задачей при этом остается представление диалектного текста в таком виде, чтобы зафиксировать его специфические особенности, отличающие его (по разным аспектам) от литературного языка, и предоставить пользователям корпуса возможность найти контексты с диалектной спецификой, не допуская при этом излишней интерпретации языкового материала. Наряду с учетом структурно-лингвистических особенностей текстов представление в XML-формате позволяет использовать для поиска в документах современные поисковые механизмы, основанные, в частности, на X-Path, которые значительно ускоряют поиск по большим объемам данных. Также по мере добавления текстов в корпус составляется словарь, который позволяет определить типичные для данной местности речевые обороты и диалектные слова, одновременно словарь ускоряет поиск в текстах. Пример текста в XML-формате приведен на рисунке 2.

```
<html>
 <head>
  <meta content="Krajnova1" name="fname"/>
...
 </head>
 <body>
  <noindex>Диалектолог: А вот раньше было, там не разрешали крест носить и вообще
веровать - как тогда было?
  </noindex>
  <speech actor="Крайнова Антонина Егоровна" sex="жен" age="60" profession="пенсионер"
sub-topic="Религия" sub_type="Рассказ-повествование">
   <w>
    <ana lex="а" gr="CONJ"/>
    <ana lex="а" gr="CONJ"/>
    а
   </w>
   <w>
    <ana lex="весь" gr="A,IDIOM=sg,n,nom"/>
    <ana lex="весь" gr="A,IDIOM=sg,n,nom"/>
    всё
   </w>
   <w>
    <ana lex="равно" gr="PRAEDIC,IDIOM"/>
    <ana lex="равно" gr="PRAEDIC,IDIOM"/>
    равно
   </w>
   <punct>//</punct>
...
   <w>
    <ana lex="пересилить" gr="V=pf,indic,praet,pl=nstand"/>
    <ana lex="преодолеть" gr="V=pf,indic,praet,pl=nstand"/>
    пересилили
   </w>
   <punct>/</punct>
   <w>
    <ana lex="дочка" gr="S,f,anim=sg,nom"/>
    <ana lex="дочка" gr="S,f,anim=sg,nom"/>
    дочка
   </w>
   <punct>//</punct>
  </speech>
...
 </body>
</html>
```

Рис. 2. Размеченный диалектный текст, конвертированный в XML-ориентированный формат

На современном этапе развития компьютерных наук и информационных технологий перед исследователем-разработчиком открыт широкий спектр программных инструментов и форматов представления данных. Однако выбор конкретных технологий, решений и алгоритмов всегда несколько ограничивается особенностями той предметной области, которая подлежит автоматической обработке. Как показано в данной работе, при создании такого нетривиального ресурса как мультимедийный диалектный корпус необходимо учитывать специфику диалектного текста, а также ориентацию на разные типы представления единицы корпуса. В настоящее время разработка электронной базы корпуса идет параллельно в двух направлениях: использование классической реляционной СУБД и XML-ориентированной СУБД.

1. Список литературы

[1] *Гольдин В.Е., Крючкова О.Ю.* Текстовый диалектологический корпус как модель традиционной сельской коммуникации // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 7 (??). М.: РГГУ, 2008. С. 268-273.

[2] *Батраева И.А., Гольдин В.Е., Крючкова О.Ю.* Поисковый механизм Саратовского диалектного корпуса // Компьютерные науки и информационные технологии. Материалы международной научной конференции 1-4 июля 2009 г. Саратов: Изд-во СГУ, 2009. С. 24-27.

[3] *Гольдин В.Е., Крючкова О.Ю.* Тематическая разметка и тематический анализ диалектного текстового корпуса // Языковая личность – текст – дискурс: теоретические и прикладные аспекты исследования. Самара, 2006. Ч.1. С. 71-80.

[4] *Поляков А.Е.* Технология подготовки информации в Национальном корпусе русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., 2005. С. 175-192.

[5] *Ляшевская О.Н., Плунгян В.А., Сичинава Д.В.* О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., 2005. С. 111-135.