Математические структуры и моделирование 2018. № 4(48). С. 85–91

УДК 004.622:801.81
DOI: 10.25513/2222-8772.2018.4.85-91

ПРЕДСТАВЛЕНИЕ РАЗМЕТКИ КОРПУСА НАРОДНОЙ РЕЧИ СРЕДНЕГО ПРИИРТЫШЬЯ

Д.Н. Лавров
к.т.н., доцент, e-mail: lavrov@omsu.ru
М.А. Харламова
к.фил.н., доцент, e-mail: khr-spb@mail.ru
Е.А. Костюшина
д.ф.-м.н., доцент, e-mail: kea.omsu@gmail.com

Омский государственный университет им. Ф.М. Достоевского, Омск, Россия

Аннотация. В статье рассматриваются способы репрезентации диалектных записей в региональном корпусе. В центре внимания — модели представления тематической, структурной и отчасти фонетической разметок. Особое внимание уделяется и модели представления экстралингвистических данных. Предложенные решения основаны на представлении реляционных баз данных и формате XML.
Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-012-00519.

Ключевые слова: тематическая разметка, метатекстовая разметка, формат XML, региональный диалектный корпус.

Введение

Корпус народной речи Среднего Прииртышья формируется за счёт сбора и последующей расшифровки записанных в экспедициях диалектных текстов. Для хранения полученных данных разрабатывается специализированная информационная система — корпус народной речи.

Ранее в рамках проекта электронного словаря была разработана система для репрезентации фонетических особенностей говоров Среднего Прииртышья [1–3]. В рамках нового проекта — регионального корпуса народной речи — перед коллективом стоят следующие задачи: (1) описать манифестацию в корпусе экстралингвистической информации; (2) описать структурную и тематическую разметки текстов.

1. Экстралингвистическая разметка

Модель экстралингвистической информации после проведённого анализа распадается на две сущности: «Паспорт информанта» и «Паспорт текста». Анализ позволил выделить атрибуты каждой сущности.

Паспорт информанта:

- Фамилия — lname.
- Имя — fname.
- Отчество — sname.
- Пол — gender.
- Год рождения — birth_year.
- Место рождения — birth_location.
- Место рождения родителей — birth_parent_location.
- Кем себя считает — who_i_am.
- Образование — education.
- Род занятий — occupation.
- Говор — dialect.

Паспорт текста:
- Место записи — location.
- Год записи — year.
- Источник (материальный носитель записи) — source.
- Размеченный текст — record.

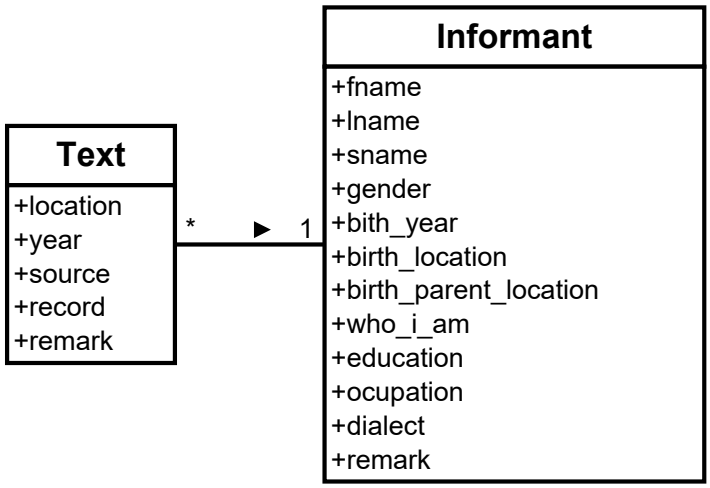

Рис. 1. Модель экстралингвистической информации

2. Фонетические знаки и их представление

В редакторе разметки необходимо ограничить количество вводимых символы для того, чтобы, с одной стороны, показывать фонетику, а с другой — не дозволять ввода служебных для XML символов. Для представления фонетических знаков принято решение использовать utf-8, а также стандартные теги и символы HTML (см. табл. 1).

Таблица 1. Представление фонетических знаков

Фонетический знак	**Внутреннее представление**	**Пример**
ў	&#1118 (или ў в utf-8)	ў-Репинки, хлеў
w	w	krawa, маркowка
h	h	hр'ибы, ahароч'чик
γ	&gamma	Мноγа
’	’	Жен'шина
”	”	ч”иста
^{a}	^a	чисто^{a}
l	l	Былo

3. Структурная разметка

Структурная разметка выполняется на основе формата XML. Для структурной разметки достаточно двух пар тегов: <вопрос>...</вопрос> и <ответ>...</ответ>.

4. Тематическая разметка

Для отображения тематической разметки предлагается использовать русскоязычные теги, название которых совпадают с названиями тем. Темы образуют иерархическую древовидную структуру. Вложения тем друг в друга описываются знаком «:». Так, если в тексте актуализируется тема «родина» и её подтема «деревня», то соответствующий тег будет <родина:деревня>.

Пример внешнего представления разметки Пример размеченного текста с фонетической, структурной и фонетической разметками:

```
<вопрос>А жили вы где? В какой деревне?</вопрос>
<ответ> Юрьифка//</ответ>
<вопрос>
   Сестра сказала, что вы последней оттуда съехали?
</вопрос>
<ответ>
   Да//
   Да/ Пач'ти последняя/
   <жизнь>
       Жалею вот ужэ пятый гот жывy кажэца-и
       жыз'нь и-жывёш
   </жизнь>
   / ни-магy привыкнуть г-гораду//
   <родина:деревня>
      Панимаити ни-магy я привыкнуть/
```

```
      а-там-эт жыла/ деревня свая//
      Природа и-кажэца вот выросла там/
      там радилась /там-и моладась мая
      прахадила/ там детей наражала/
      ну-вот фсё идиал'на// А-время-та
      нашэ како идиал'на-та была //
   </родина:деревня>
</ответ>
```

Обратите внимание на то, что данное представление используется только для отображения на экране в редакторе разметки (так что использование символа «:» на данном этапе некритично), внутреннее представление иное, и о нём пойдёт речь в следующем разделе.

5. Внутренне представление, используемое для обмена данными между приложениями

Предыдущий раздел описывал внешнее представление разметки.

Для визуализации разметки чем короче тег, тем лучше. Это вполне устраивает и разработчиков, и программистов. Казалось бы, почему не использовать это представление и для обмена данными между приложениями?

Есть несколько причин. На уровне спецификаций без внешних описаний только по названию тега невозможно определить, какой это тег — структурный или тематический. Кроме того, в указанном выше представлении вложение тем обозначается двоеточием, что для форматов HTML и XML неприемлемо. В тоже время исследователи-филологи активно его используют при выполнении ручной разметки. Решение состоит в создании внутреннего представления данных, которое будет скрыто от пользователя приложения-редактора.

Принципы, реализованные во внутреннем представлении.

- Все названия тегов — и структурных, и тематических — заменены на английские названия.
- Экстралингвистическая разметка соответствует полям таблиц базы данных (см. рис. 1).
- Структурные теги превращаются в `<question>` и `<answer>`.
- Тематический тег один `<theme class="тема--подтема">`.

Пример представления экстралингвистической, тематической и фонетической разметок во внутреннем формате для обмена данными между разрабатываемыми приложениями (данные вымышленные).

```
<doc>
  <informant>
    <fname>Ольга</fname>
    <sname>Карловна</sname>
    <lname>Карнелс</lname>
    <gender>женский</gender>.
    <birth_year>1930</birth_year>
```

```
  <birth_location>
     д. Новоникольск, жила в д. Баженово
     Тарского района 10 лет
  </birth_location>
  <birth_parent_location>
     родители переехали из Белоруссии, д. Николка
     в 1961 г. в Большие Уки
  </birth_parent_location>
  <who_i_am>
     считает себя «российской»
  </who_i_am>
  <education>4 класса</education>
  <ocupation>
      пенсионерка, сортировщик на почте
  </ocupation>
  <dialect>старожильческий</dialect>
  <remark>
     Год прожила в Казахстане, около 30 лет прожила
     в Таджикистане, 9 лет жила в Новосибирске.
  </remark>
</informant>

<text>
  <location>
     д. Большие Уки Большеуковский район
  </location>
  <year>2005</year>
  <source>
     тетрадь №122, кассета №82,
     записи произведены: Митюшовой Ириной,
     гр. ЯФ - 303, Полозковой Марией, гр. ЯФ – 302.
  </source>
  <remark></remark>
  <record>
    <![CDATA[
    <question>А жили вы где? В какой деревне?</question>
    <answer><b>Ю</b>рьифка//</answer>
    <question>
       Сестра сказала, что вы последней оттуда съехали?
    </question>
    <answer>
      Д<b>а</b>//
      Д<b>а</b>/ Пач’т<b>и</b> посл<b>е</b>дняя/
      <theme class="жизнь">
        Жал<b>е</b>ю вот уж<b>э</b> пятый г<b>о</b>т
        жыв<b>у</b> к<b>а</b>жэца-и ж<b>ы</b>з’нь
        и-жыв<b>ё</b>ш
      </theme>
```

```
        / ни-маг<b>у</b> прив<b>ы</b>кнуть
        г-г<b>о</b>раду//
        <theme class="родина--деревня">
          Паним<b>а</b>ити ни-маг<b>у</b> <b>я</b>
          прив<b>ы</b>кнуть/ а-т<b>а</b>м-эт
          жыл<b>а</b>/ дир<b>е</b>вня сва<b>я</b>//
          Прир<b>о</b>да и-к<b>а</b>жэца в<b>о</b>т
          в<b>ы</b>расла т<b>а</b>м/ т<b>а</b>м
          радил<b>а</b>сь /т<b>а</b>м-и м<b>о</b>ладась
          ма<b>я</b> прахад<b>и</b>ла/ там дит<b>е</b>й
          нараж<b>а</b>ла/ ну-в<b>о</b>т ӧсё
          иди<b>а</b>л'на// А-вр<b>е</b>мя-та н<b>а</b>шэ
          как<b>о</b> иди<b>а</b>л'на-та б<b>ы</b>ла //
        </theme>
      </answer>
      ]]>
    </record>
  </text>
</doc>
```

Использование CDATA позволяет не заботиться о точном соответствии спецификациям XML внутри поля record и без дополнительных преобразований использовать данный код для отображения на HTML-странице web-приложения.

Заключение

В настоящее время на основе разработанной модели представления созданы два прототипа приложений: десктоп-редактор для создания разметки в условиях экспедиций и отсутствия доступа к среде интернет и веб-приложение, позволяющее делать выборку из базы данных на основе MySQL по экстралингвистической информации и отображать её в виде HTML-страниц с возможностью интерактивной тематической разметки. Прототип веб-приложения создан на языке Python с использованием фреймворка Django и библиотеки jQuery.

В момент написания статьи проходило опытное тестирование и апробация указанных прототипов.

Результаты данной статьи были представлены в докладе на конференции «Математическое и компьютерное моделирование» [4].

Благодарности

Выражаем признательность Лапину Александру Петровичу и Черкащенко Илье Александровичу за ценные замечания и помощь в реализации прототипов. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №18-012-00519.

Литература

1. Лавров Д.Н., Харламова М.А. Словарь констант народной речи: выбор платформы представления // Вестник Омского университета. 2015. № 1(75). С. 213–216.
2. Харламова М.А. Константы народной речемысли и их лексикографическая интерпретация. Омск : Изд-во Ом. гос. ун-та, 2014. 290 с.
3. Балезин И.А., Лавров Д.Н., Харламова М.А. Архитектура мобильного клиента под iOS для доступа к веб-словарю народной речи Среднего Прииртышья // Математические структуры и моделирование. 2016. № 4(40). С. 133–142.
4. Лавров Д.Н., Харламова М.А., Костюшина Е.А. Модель представления экстралингвистической и тематической разметки в корпусе народной речи // VI-я Междунар. науч. конф. «Математическое и компьютерное моделирование», посвящ. памяти проф. Б.А. Рогозина. 23 ноября 2018. С. 115–118.

REPRESENTATION OF THE CORPUS OF MEDIUM IRTYSH FOLK DIALECT

D.N. Lavrov
Ph.D. (Eng.), Associate Professor, e-mail: lavrov@a.ru
M.A. Kharlamova
Ph.D. (Philological), Associate Professor, e-mail: khr-spb@mail.ru
E.A. Kostushina
Ph.D. (Eng.), Associate Professor, e-mail: kea.omsu@gmail.com

Dosotevsky Omsk State University, Omsk, Russia

Abstract. The article discusses ways of representing dialect entries in the regional corpus. The focus is on models for the presentation of thematic, structural and partly phonetic markings. Particular attention is paid to the presentation model of extralinguistic data. The proposed solutions are based on the representation of relational databases and XML format.
The reported study was funded by RFBR according to the research project № 18-012-00519.

Keywords: thematic markup, metatext markup, XML format, regional dialect body.

Дата поступления в редакцию: 20.11.2018