Математические структуры и моделирование 2019. № 3(51). С. 135–144

УДК 004.622:801.81
DOI 10.25513/2222-8772.2019.3.135-144

ОБРАБОТКА ТЕКСТА С ФОНЕТИЧЕСКОЙ И ТЕМАТИЧЕСКОЙ РАЗМЕТКАМИ ДЛЯ ОТОБРАЖЕНИЯ ТЕКСТОВ ДИАЛЕКТНОГО КОРПУСА

Д.Н. Лавров
к.т.н., доцент, e-mail: lavrov@omsu.ru
А.П. Лапин
студент, e-mail: aleksandrlapinsanek@gmail.com
М.А. Харламова
к.фил.н., доцент, e-mail: khr-spb@mail.ru
И.А. Черкащенко
магистрант, e-mail: ilyachr@yahoo.com

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

Аннотация. В работе описаны схемы конечных автоматов, положенные в основу разработки веб-приложения диалектного корпуса Среднего Прииртышья. В статье предложена доработка структуры XML-документа, представляющего текстовую запись интервью с одним или несколькими информантами с фонетической и тематической разметками. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-012-00519.

Ключевые слова: веб-разработка, тематическая разметка, конечные автоматы, диалектный корпус.

Введение

На протяжении более чем полувека в диалектологических экспедициях были собраны материалы в виде интервью с носителями говоров с разной диалектной основой. Все полученные данные хранятся в аудио- и видеоформате. Аудиоформат анализируется лингвистами-диалектологами, а затем преобразуется в текстовую форму с фонетической разметкой. Расшифрованный текст с фонетической разметкой размечается по тематике содержания. Совокупность всех размеченных текстов с соответствующими аудио- и видеозаписями представляет собой диалектный корпус полиэтнического региона Среднего Прииртышья, над созданием которого мы работаем.

В серии предыдущих работ [1–3] был создан формат представления тематической разметки. В статье представлена доработанная версия этого формата и описание машины конечных состояний по перекодированию этих данных в представление для веб-страниц. Такой подход позволяет хранить размеченные тексты в стандартной реляционной базе данных (БД), а при поступлении запроса отображать данные с возможностью подсветки заданных тем и с отображе-

нием реквизитов текстов и паспортов информантов. Использование библиотек jquery позволяет выделять темы, не перегружая страницу. Аналогичный проект [4] может выделять только одну тему, а саму подсветку темы в тексте может выполнять только лишь с перезагрузкой страницы.

Целью данной работы является построение приложения для отображения тематической и фонетической разметок в веб-приложении.

1. Структура входного документа

Как правило, диалектные записи делаются в виде интервью с одним информантом. В ходе лингвистического анализа текстов было обнаружено, что часть интервью является записью беседы в форме диалога, реже — полилога.

В ранее разработанную структуру XML-документа [2] было добавлено поле, идентифицирующее информанта, что позволило использовать этот идентификатор для обозначения респондента в тексте.

Были добавлены соответствующие изменения в разрабатываемый код приложения разметки [2]. Внешний вид упомянутого приложения показан на рисунке 1.

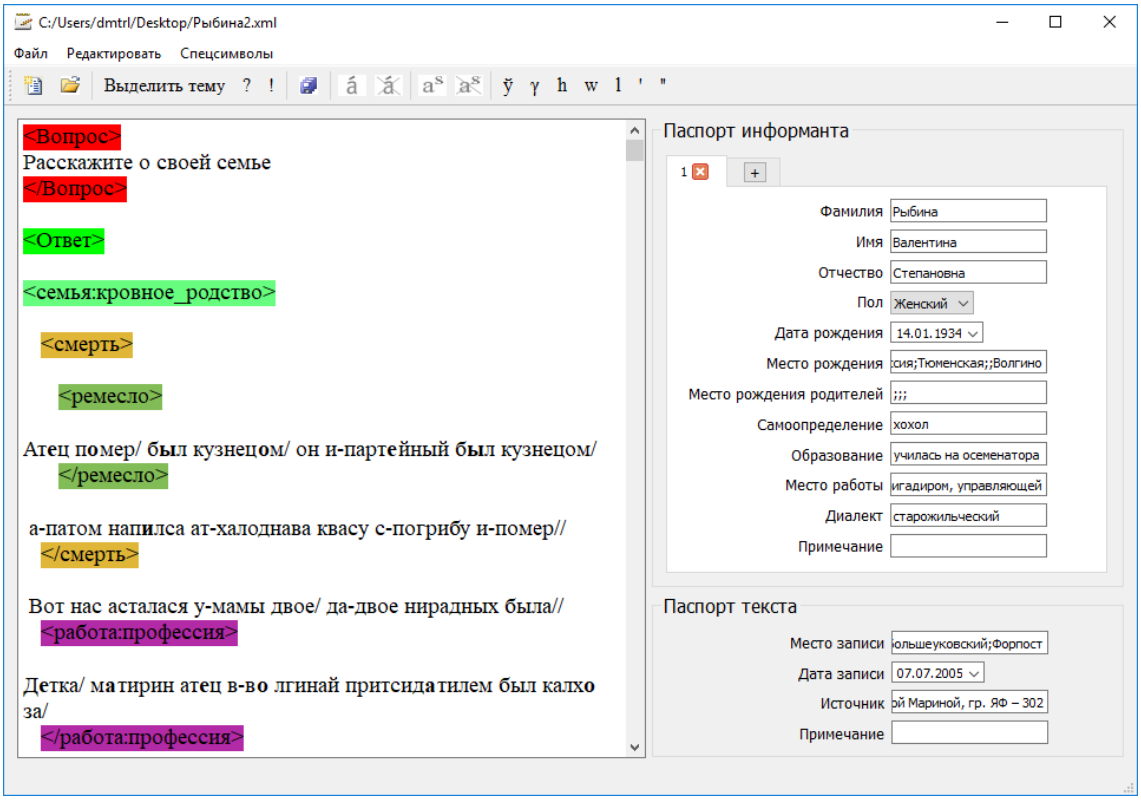

Рис. 1. Внешний вид приложения для выполнения тематической разметки

Необходимо, чтобы на веб-сайте регионального корпуса можно было «подсветить» темы, присутствующие в размеченных текстах. Для этого в текст добавлены теги `<theme>` со свойством `class`, в котором прописывается название

темы. Пример: `<theme class="Природа">`. Теги могут быть сложными, когда одна тема является подтемой другой темы. В таких темах используется разделитель «двойной минус» `--`. Пример: `<theme class="Семья--Брат">`. Окончание темы обозначается закрывающим тегом `</theme>`. Фрагменты текста с разной тематической разметкой могут быть вложены друг в друга, но с обязательным соблюдением корректной скобочной структуры: все внутренние теги должны быть закрыты до закрытия тега контейнера.

Текст интервью состоит только из вопросов и ответов. Для того, чтобы выделить вопросы и ответы, на них в текст добавлены теги `<question>` — вопрос и `<answer>` — ответ. Если информантов несколько, то к тегу `<answer>` добавляется свойство `<informant>`, в котором прописывается идентификатор интервьюируемого. В тексте могут присутствовать символы кириллицы, латиницы, а также специальные символы, обозначающие те или иные звуки. В тексте могут использоваться такие html-теги, как <b> — для выделения ударений или <sub> — для выделения особенностей звучания.

Пример кода разметки с идентификаторами нескольких информантов приведён ниже:

```
<doc>
 <informant>
  <id>1</id>
  <lname>Сидоров</lname>
  <fname>Илья</fname>
  <sname>Николаевич</sname>
 </informant>
 <informant>
  <id>2</id>
  <lname>Сидорова</lname>
  <fname>Анастасия</fname>
  <sname>Степановна</sname>
 </informant>
 <text>
  <location>
   <state>Россия</state>
   <locality>Омск</locality>
  </location>
  <year>2019</year>
  <record><![CDATA[
  <question>
     Здравствуйте. Представьтесь, пожалуйста.
  </question>
  <answer informant="1">
       Здравствуйте. Меня зовут Илья
       Никол<b>а</b>евич.
  </answer>
  <answer informant="2">
       Здравствуйте. Мое имя Анастасия
```

```
      Степановна.
      Мне 34 года и я работаю бухгалтером.
  </answer>
  ]]></record>
 </text>
</doc>
```

2. Модель базы данных

База данных для хранения диалектных тестов с фонетической и тематической разметками содержит две таблицы «Текст» и «Информант» [1]. У текста будут следующие атрибуты: идентификатор текста, место записи (состоит из страны, региона, района и населённого пункта), год, ссылка на бумажный оригинал (запись на магнитную ленту или другой носитель), заметки и сам текст. У таблицы «Информант» имеются следующие поля: уникальный идентификатор, фамилия, имя, отчество, пол, место рождения, место рождения родителей, кем себя считает, образование, род занятий, тип диалекта и заметки.

Каждый текст должен быть прикреплён как минимум к одному информанту, а информант связан с одним текстом. Это связь «многое-ко-многим». Данные на сайт передаются в виде отдельных документов, в каждом из которых хранится вся информация о расшифрованном тексте, а также вся информация о респондентах данной записи.

В период разметки текста отсутвуют данные об уникальных идентификаторах информантов. А значит, нельзя сопоставить с достоверностью информантов из поступившего документа и из базы данных. Из-за этого принято решение каждого нового информанта хранить отдельно, даже если он уже есть в базе данных. Это нарушение первой нормальной формы, но это необходимо в условиях данной задачи. Чтобы это реализовать, была изменена структура базы данных: связь «Информант–Текст» реализована как «многие к одному». Для её реализации к паспорту информанта добавлен идентификатор конкретного текста.

3. Модель передачи данных

Расшифрованные тексты в базу данных передаются в виде xml-документа, структура которого была описана выше, и соответствуют в целом структуре базы данных. В документе существует главный тег `<doc>`. Внутри него может быть тег `<version>`, чтобы в случае изменений можно было легко отличить старую структуру документа от нового, а также обязательно есть как минимум один тег `<informant>` и один тег `<text>`. Тег `<informant>` может содержать те же теги, что и поля соответствующих таблиц в базе данных. Если информантов несколько, то они должны содержать тег `<id>`, на который необходимо ссылаться в тексте, в свойстве `<informant> тега \verb`<answer>". Если какой-либо тег отсутствует, то атрибут в базе данных остаётся пустым.

Атрибут, в котором содержится текст статьи, имеет секцию «CDATA». Сделано это для разделения точно определённых структур описания информанта и параметров интервью от тематической разметки самого интервью, которое в свою очередь может не полностью соответствовать всем спецификациям XML. Чтобы разобрать данный документ и внести полученную информацию в БД, будет использоваться библиотека «lxml». Эта библиотека разбирает документ только один раз при его загрузке.

В файле текст форматирован удобным для прочтения xml образом. При отображении на сайте необходимо заменить тематические теги, а также теги вопросов и ответов на теги html. Ниже приведён пример такого преобразования.

Текст, хранящийся в базе данных:

```
<question>Здравствуйте. Представьтесь, пожалуйста.</question>
<answer informant="1">
    Здравствуйте. Меня зовут Иль<b>я</b> Никол<b>а</b>евич.
</answer>
<answer informant="2">
    Здравствуйте. Мое имя Анастас<b>и</b>я Степ<b>а</b>новна.
    Мне 34 года и я работаю бухгалтером.
</answer>
<question>Расскажите что-нибудь с своей семье.</question>
<answer informant="1">
   <theme class="Семья">
     <theme class="Семья--Образование">
          Моя жена, в отличии от меня имеет высшее образование.
     </theme>
     <theme class="Работа">
          Благодаря ему она смогла устроиться бухгалтером
          в небольшую компанию. Ей очень нравиться там
          работать. Я же работаю на СТО и, к моему удивлению,
          зарабатываю больше
     </theme>
   </theme>
</answer>
```

для отображения на веб-странице преобразуется в:

```
<ins class="question">
<i>Вопрос: </i>Здравствуйте. Представьтесь, пожалуйста.
</ins>
<ins class="answer">
    <i>И. Н.: </i>Здравствуйте. Меня зовут Иль<b>я</b>
    Никол<b>а</b>евич.
</ins>
<ins class="answer">
    <i>А. С.: </i>Здравствуйте. Мое имя Анастас<b>и</b>я
    Степ<b>а</b>новна. Мне 34 года и я работаю
```

```
    бухгалтером.
</ins>
<ins class="question">
   <i>Вопрос: </i>Расскажите что-нибудь с своей семье.
</ins>
<ins class="answer">
   <i>И. Н.: </i>
      <ins class="my_tag_0 ">
          <ins class="my_tag_0 my_tag_0 my_tag_1 ">
             Моя жена, в отличии от меня
             имеет высшее образование.
      </ins>
      <ins class="my_tag_0 my_tag_2 ">
         Благодаря ему она смогла устроиться бухгалтером
         в небольшую  компанию. Ей очень нравиться там
         работать. Я же работаю на СТО и, к моему
         удивлению, зарабатываю больше.
      </ins>
   </ins>
</ins>
</div>
```

После такого преобразования подсветка тем осуществляется через изменение стилей «на лету» с использованием библиотеки `jquery`. Это не требует перезагрузки страницы и улучшает удобство использования приложения.

Первый конечный автомат удаляет лишние пробельные символы (см. рисунок 2). После разбора документа и нахождения в нём размеченного текста интервью запускается первый автомат. Он удаляет лишние пробелы из текста. При нахождении нескольких пробельных символов, идущих подряд, он оставляет только первый, удаляя все остальные. Этот автомат удаляет пробельные символы после открывающихся и перед закрывающимися тегами `<question>` и `<answer>`. Кроме того, до и после открывающегося и закрывающегося тега `<theme>` на вход автомату поступают локальные идентификаторы информантов, используемые в файле, и соответствующие им глобальные идентификаторы информантов, используемые в базе данных. Автомат заменяет одни на другие. Если в тексте был обнаружен идентификатор информанта, которого нет в документе, то новый идентификатор будет равен «−1».

Второй автомат работает при открытии страницы с интервью. Автомат заменяет открывающийся тег `<question>` на текст `<ins class="question"><i>Вопрос: </i>"`. А открывающийся тег `<answer>` на аналогичный текст `<ins class="answer"><i>Ответ: </i>`. При наличии в теге свойства `<informant>` вместо слова `<Ответ>` будут написаны инициалы соответствующего информанта или слово «Неизвестный», если он будет равен «−1». Открывающий тег `<theme>` меняется на тег `<ins>`. Если значение свойства `<class>` тематического тега встречается впервые, то вместо него прописывается уникальный идентификатор с использованием

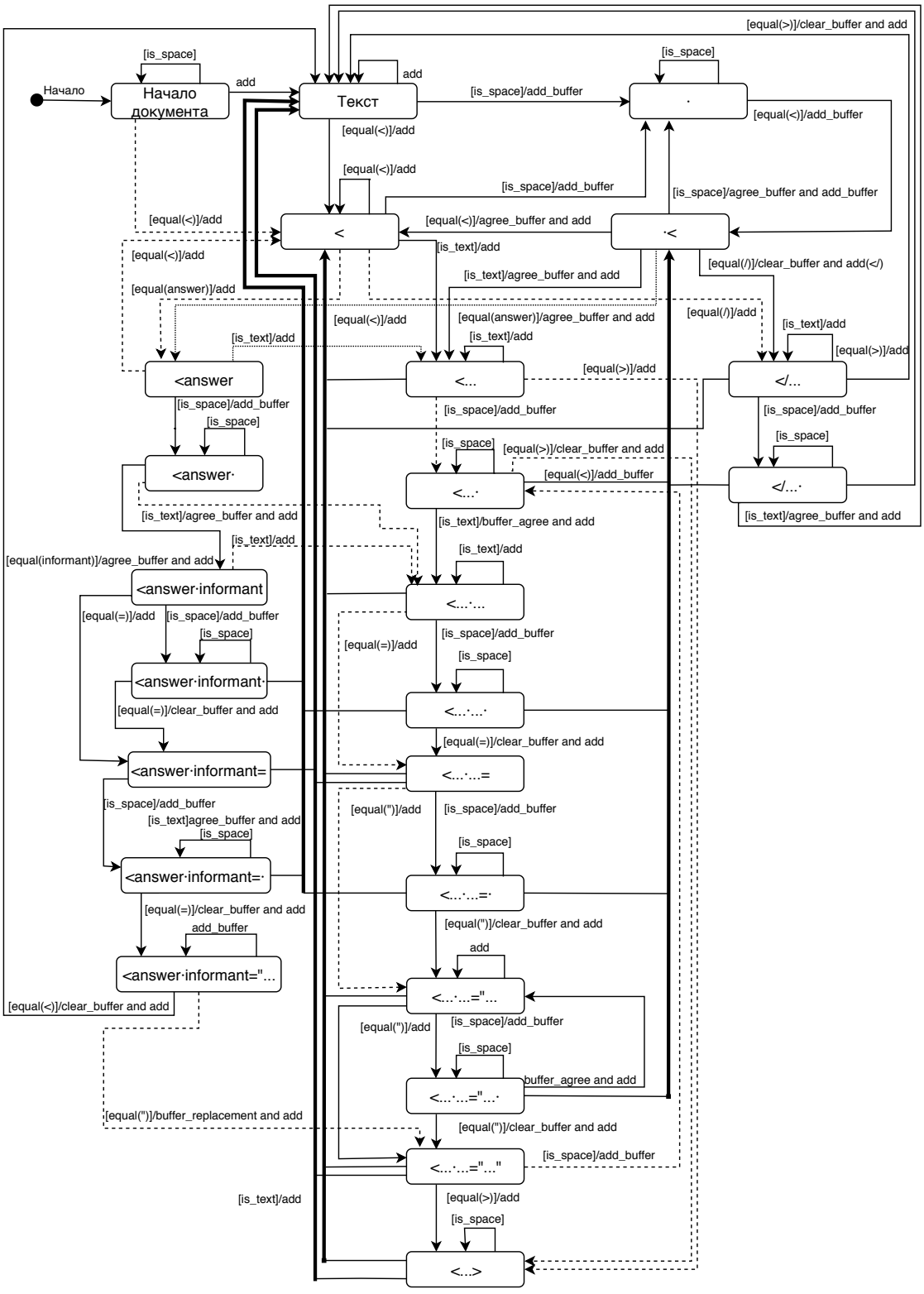

Рис. 2. Структурная схема первого автомата

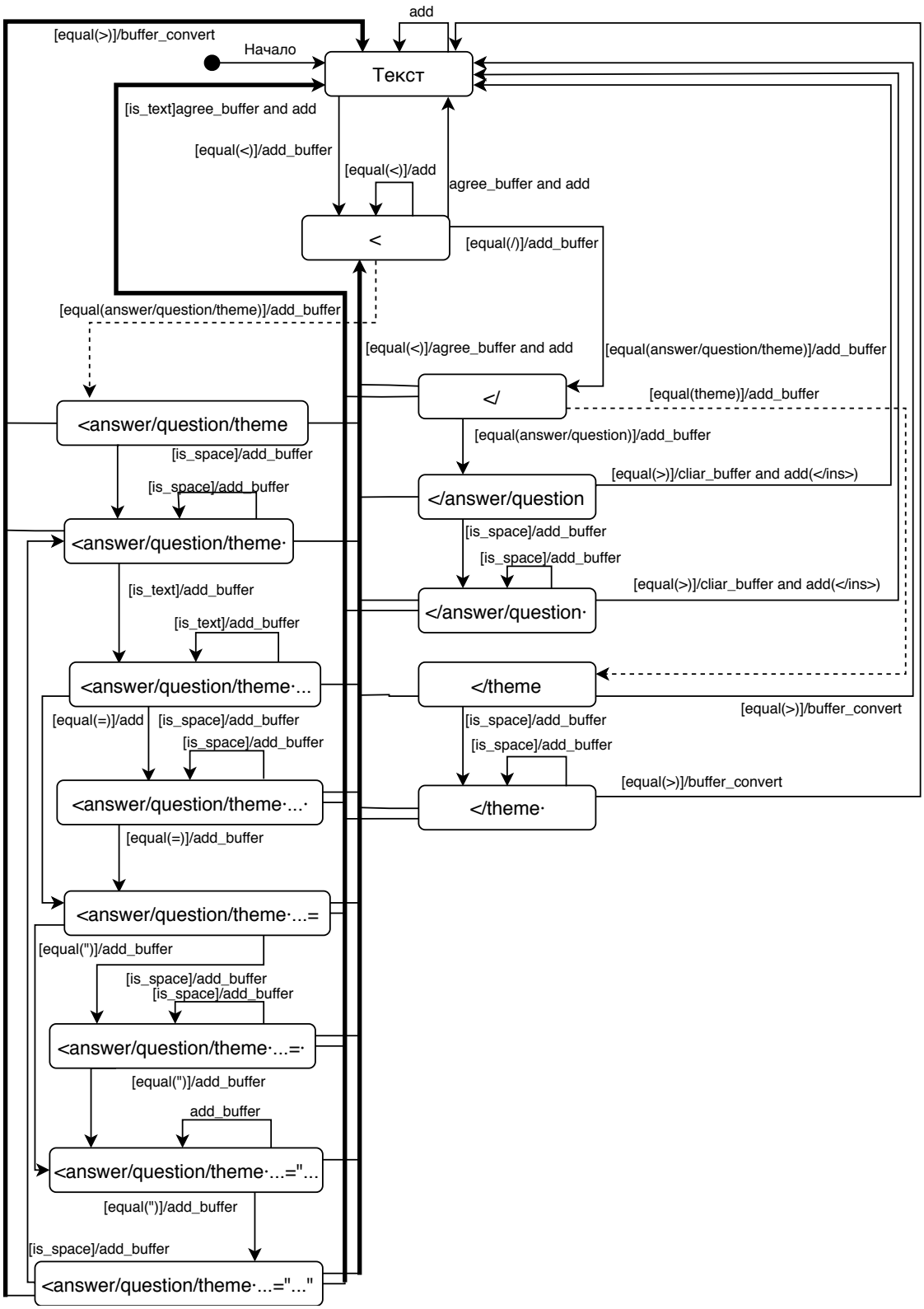

Рис. 3. Структурная схема третьего автомата

только латинских символов, цифр и нижнего подчёркивания, который и прописывается в свойстве `<class>`. Если значение тега уже встречалось, то ему присваивается тот же самый идентификатор, что и в прошлый раз. Если тематический тег является тегом темы (в иерархии тем), (он пишется через двойной минус «--»), обрабатывается как новая тема и записывается через пробел. После завершения работы автомат возвращает список всех названий тем и присвоенных им уникальных идентификаторов.

Закрывающий тег `<theme>` меняется на закрывающий тег `<ins>`. Тематический тег, установленый внутри другого тега, меняет своё значение на соответствующий идентификатор. Кроме того, прописывается идентификатор темы родительского тега. Стандартные теги html, используемые в тексте, остаются без изменений. Структурная схема автомата представлена на рисунке 3.

Заключение

В результате: (1) разработано веб-приложение на платформе Django для внесения в базу данных и отображения диалектных текстов Среднего Прииртышья, (2) доработана структура формата обмена данными между подсистемами.

В настоящее время ведётся обучение пользователей и получение обратной связи об удобстве использования и необходимости доработки программных продуктов проекта построения регионального корпуса говоров Среднего Прииртышья.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-012-00519.

Литература

1. Харламова М.А., Лавров Д.Н. История полиэтнического региона в «зеркале» народной речи: о проекте регионального диалектного корпуса. / Current trends and Future Perspectives in Russian Studies // Proceedings of the International Conference on Russian Studies at the University of Barcelona, MKR-Barcelona. Barselona : Trialba Ediciones 2018. P. 1564–1572.
2. Лавров Д.Н., Харламова М.А., Костюшина Е.А. Представление разметки корпуса народной речи Среднего Прииртышья // Математические структуры и моделирование. 2018. № 4(48). С. 85–91.
3. Лавров Д.Н., Харламова М.А., Костюшина Е.А. Модель представления экстралингвистической и тематической разметки в корпусе народной речи // Математическое и компьютерное моделирование : сборник материалов VI Международной научной конференции, посвящённой памяти Б.А. Рогозина (Омск, 23 ноября 2018 г.). Омск : Изд-во Ом. гос. ун-та, 2018. С. 115–118.
4. Лаборатория общей и сибирской лексикографии. Режим доступа: URL: `http://losl.tsu.ru/?q=corpus/demo` (дата обращения: 21.05.19).

TEXT PROCESSING WITH PHONETIC AND THEMATIC MARKUP FOR DISPLAYING TEXTS OF DIALECT CORPUS

D.N. Lavrov
Ph.D.(Eng.), Associate Professor, e-mail: lavrov@omsu.ru
A.P. Lapin
Student, e-mail: aleksandrlapinsanek@gmail.com
M.A. Kharlamova
Ph.D.(Phil.), Associate Professor, e-mail: khr-spb@mail.ru
I.A. Cherkashchenko
Master, e-mail: ilyachr@yahoo.com

Dostoevsky Omsk State University, Omsk, Russia

Abstract. The paper describes the developed finite state machine schemes that underlie the development of a web application for the dialect corpus of the Middle Irtysh. The paper proposes refinement of the structure of an XML document representing a text recording of an interview with one or more informants with phonetic and thematic markings. The study was carried out with the financial support of the Russian Federal Property Fund in the framework of the scientific project No. 18-012-00519.

Keywords: web development, thematic markup, state machines, dialect corpus.

References

1. Kharlamova M.A. and Lavrov D.N. Istoriya polietnicheskogo regiona v ”zerkale” narodnoi rechi: o proekte regional’nogo dialektnogo korpusa. Current trends and Future Perspectives in Russian Studies, Proceedings of the International Conference on Russian Studies at the University of Barcelona, MKR-Barcelona, Barselona, Trialba Ediciones 2018, pp. 1564–1572. (in Russian)
2. Lavrov D.N., Kharlamova M.A., and Kostyushina E.A. Predstavlenie razmetki korpusa narodnoi rechi Srednego Priirtysh’ya. Matematicheskie struktury i modelirovanie, 2018, no. 4(48), pp. 85–91. (in Russian)
3. Lavrov D.N., Kharlamova M.A., and Kostyushina E.A. Model’ predstavleniya ekstralingvisticheskoi i tematicheskoi razmetki v korpuse narodnoi rechi. Matematicheskoe i komp’yuternoe modelirovanie: sbornik materialov VI Mezhdunarodnoi nauchnoi konferentsii, posvyashchennoi pamyati B.A. Rogozina (Omsk, 23 noyabrya 2018 g.), Omsk, Izd-vo Om. gos. un-ta, 2018, pp. 115–118. (in Russian)
4. Laboratoriya obshchei i sibirskoi leksikografii. Rezhim dostupa: URL: `http://losl.tsu.ru/?q=corpus/demo` (21.05.19). (in Russian)

Дата поступления в редакцию: 26.08.2019