ОПИСАНИЕ СИНТАКСИЧЕСКОГО И МОРФОЛОГИЧЕСКОГО РАЗБОРА ТЕКСТА С ИСПОЛЬЗОВАНИЕМ XML

Суровцова Т. Г.

(Россия, Пущино)

Предложен формат для описания синтаксических и морфологических разборов текстов в соответствии с грамматикой русского языка с использованием языка разметки XML. Для формата создано описание DTD, руководствуясь которым можно проводить разбор текста. Формат предложен для публичного использования.

Введение. Анализ существующих и опыт разработки собственных программных продуктов для поддержки и проведения лингвистических исследований, показал, что существуют проблемы, связанные с повторной обработкой уже полученных данных. Материал, который накапливается в одной программной системе сложно, а иногда невозможно, обработать в другой. В случае, когда синтаксические или морфологические разборы текстов выполнены вручную специалистами-филологами, такие потери являются неоправданными.

Для автоматизированных систем похожая ситуация возникает из-за быстрой смены поколений вычислительной техники, морального устаревания программ, работа которых не поддерживается новыми операционными системами, вплоть до того, что форматы, в которых хранились данные часто уже невозможно прочитать. Возникают проблемы с кодировкой кириллицы и текстов на нескольких языках, а также с системой кодирования, которая была использована для хранения данных.

Наиболее часто для представления информации в компьютере используются реляционные базы данных. Это позволяет соз-

дать структуру, которая лишена противоречивости и избыточности, полна, но в то же самое время, предполагает разбиение на ряд отношений — таблиц, совокупность которых и описывает обрабатываемые системой данные. При этом предполагается использование ключей — кодов, которые позволяют связать таблицы. Если материалы, полученные в системе, надо обработать повторно, то сначала приходится тратить временные ресурсы на то, чтобы разобраться в используемой разработчиками системе кодирования и той совокупности таблиц, которая была использована для хранения информации. Иногда используют достаточно замысловатые системы кодировки, что обычно связано с особенностями программной реализации системы.

Полученные в разных программных средах разборы текстов сложно привести к общему виду, что затрудняет объединение подготовленных материалов. Это приводит к необходимости проведения повторных разборов, дополнительной трате времени, неэффективному использованию ресурсов.

С целью унификации данных был разработан формат — описание, который позволяет хранить информацию о синтаксическом и морфологическом разборе текста в виде понятном как человеку, так и большинству программных систем, размечая «плоские» тексты с использованием языка разметки XML (eXtensible Markup Language) [4]. Результатом является DTD (Document Type Definitions), в котором описаны правила, в соответствии с которыми производятся эти разборы.

Использование XML — это универсальный способ разметки любого текста, причем описывается не содержание текста, а его структура, поясняющая смысл составляющих его компонентов, что в свою очередь облегчает дальнейшую обработку.

Материал для исследования. В Петрозаводском государственном университете ведется разработка информационной системы «Статистические методы анализа литературного текста» (ИС «СМАЛТ») [1, 3]. (Проекты РГНФ № 02-04-12015в, № 05-04-12418в, рук. Рогов А. А., http://smalt.karelia.ru), которая в настоя-

щее время содержит ряд литературных произведений (публицистические статьи из журналов «Время», «Эпоха», «Современник», «Гражданин» и др.), их морфологические и синтаксические параметры.

Так как разработка ведется уже несколько лет, то есть разборы произведений, которые были выполнены с использованием компьютеров Macintosh фирмы Apple, а часть данных хранится во внутреннем формате ИС «СМАЛТ» в виде текстовых файлов. Для них существует возможность загрузки в файлы сервера баз данных Interbase. Блок анализа авторства произведения реализуется с использованием сервера баз данных Oracle, то есть даже в рамках одной системы существуют данные, представленные в разных форматах.

Разборы текстов выполнены с использованием дерева синтаксического и морфологического разбора, которое создано филологами кафедры русского языка Петрозаводского государственного университета. Это дерево и легло в основу разработанного формата на основе языка XML, в который включены синтаксический и морфологический разборы для текстов на русском языке, выполненные в соответствии с русской грамматикой [2].

Разработка формата описания морфологических и синтаксических разборов. В качестве основы для описания разметки с использованием языка XML был выбран формат DTD, который имеет большую историю, чем XMLS-схемы, которые активно разрабатываются в настоящее время (http://www.w3c.org).

Каждая статья рассматривается как набор элементов — объектов, которые связаны между собой иерархическими отношениями, каждый объект описывается набором свойств, которые их характеризуют.

Таким образом, полученный формат представляет собой набор, в котором объявлены элементы, использованные при разборе, порядок их следования и вложенность друг в друга, а также атрибуты, их описывающие.

Так было предложено ввести элемент <Статья>, который включает в себя множество атрибутов, описывающих статью, и набор предложений. Элемент <Предложение> в свою очередь содержит информацию о разборе, являясь кортежем слов и знаков препинания. Для каждого слова, в свою очередь, определяется набор атрибутов.

Набор атрибутов для элемента <Предложение> зависит от его вида, для элемента <Слово> — от части речи. Формат разбит на файлы, каждый из которых описывает некоторое логически связанное подмножество. Например, посвящен описанию атрибутов существительного или разбору сложного предложения, что удобно, так как позволяет эти части формата использовать независимо от остальных.

В описании выделено 23 вида слов, к одному из которых можно отнести любое слово, и 8 видов предложений.

Корневой элемент <Статья> определяет, что в него могут быть включены элементы, которые содержат описание самой статьи, а именно: <Название статьи>, <Автор>, <Журнал>, <Год>, <Номер журнала>, а также элементы <Предложение> и <Знак_препинания>.

Ниже приведен фрагмент описания элемента <Статья> и разбор части речи на примере существительного.

```
<!-- Начало описания элемента Статья -->
<?xml version="1.0" encoding="UTF-8"?>
<!-- SMR  XML DTD v1.0.....-->
<!-- File smr.dtd .....-->
<!-- Для публичной идентификации используется следующая строка -->
<!-- <!DOCTYPE srm PUBLIC "-//PetrSU//SRM XML DTD V1.0//RU"  " http://nlcom.onego.ru/smalt/dtd/smr.dtd">-->
<!ELEMENT Статья (Название_статьи, Автор, Журнал, Год, Номер_журнала, Предложение?)>
<!-- Название статьи может содержать любые текстовые данные -->
<!ELEMENT Название_статьи (#PCDATA)>
```

<!-- Элемент Автор статьи может содержать элементы Автор_известен, Автор_не_известен, Dubia (спорное авторство) -->

<!ELEMENT Автор (Автор_известен | Автор_не_известен | Dubia)>

<!-- Если автор или авторы известны, то перечисляются все фамилии авторов -->

<!ELEMENT Автор_известен (Имя_автора)+>

<!ELEMENT Имя_автора (#PCDATA)>

<!-- Для указания неизвестного авторства используется элемент Автор_не_известен, если статья имеет спорное авторство, то возможные автор или авторы перечисляются -->

<!ELEMENT Автор_неизвестен EMPTY>

<!ELEMENT Dubia (Имя_автора)+>

<!-- Определяются элементы, описывающие статью -->

<!ELEMENT Журнал (#PCDATA)>

<!ELEMENT Год (#PCDATA)>

<!ELEMENT Номер_журнала (#PCDATA)>

<!-- Определяем элемент Предложение, как кортеж элементов Слово и Знак_препинания -->

<!ELEMENT Предложение (Тип_предложения?, (Слово | Знак_препинания)*)>

<!-- Пропущено описание элементов Тип_предложения, Слово, Знак_препинания -->

<!-- … -->

<!-- Конец описания элемента Статья -->

Рассмотрим разбор части речи на примере существительного.

<!-- Начало описания разбора существительного -->

<?xml version="1.0" encoding="UTF-8"?>

<!-- SMR XML DTD v1.0.....-->

<!-- File suchestvitelnoe.dtd .....-->

<!—Параметры разбора существительного -->

<!ELEMENT Существительное (Начальная_форма, Разряд_по_значению_А, Разряд_по_значению_Б, Раз-

ряд_по_значению_В, Категория_рода, Категория_числа, Категория_падежа, Типы_склонения)>

<!ELEMENT Начальная_форма (#PCDATA)>

<!ELEMENT Разряд_по_значению_А EMPTY>

<!ATTLIST Разряд_по_значению_А value (Неодушевленное | Неодушевленное_в_значении_одушевленного | Одушевленное | Одушевленное_в_значении_неодушевленного) #REQUIRED>

<!ELEMENT Разряд_по_значению_Б EMPTY>

<!ATTLIST Разряд_по_значению_Б value (Нарицательное | Нарицательное_в_значении_собственного | Собственное | Собственное_в_значении_нарицательного) #REQUIRED>

<!ELEMENT Разряд_по_значению_В EMPTY>

<!ATTLIST Разряд_по_значению_В value (Абстрактное | Абстрактное_в_значении_конкретного | Вещественное | Конкретное | Конкретное_в_значении_абстрактного | Собирательное) #REQUIRED>

<!ELEMENT Категория_рода EMPTY>

<!ATTLIST Категория_рода value (Женский | Мужской | Не_имеет_рода | Общий | Средний) #REQUIRED>

<!ELEMENT Категория_числа EMPTY>

<!ATTLIST Категория_числа value (pluralia_tantum | singularia_tantum | Единственное | Множественное) #REQUIRED>

<!ELEMENT Категория_падежа EMPTY>

<!ATTLIST Категория_падежа value (Именительный | Родительный | Дательный | Винительный | Творительный | Предложный) #REQUIRED>

<!ELEMENT Типы_склонения EMPTY>

<!ATTLIST Типы_склонения value (I_склонение | II_склонение| III_склонение | Разносклоняемое | Адъективное | Несклоняемое | Нет) #REQUIRED>

<!-- Конец описания разбора существительного -->

Полученная спецификация является достаточно объемной, разбита на 19 файлов, поэтому привести ее полностью невозможно, полное описание представлено: http://nlcom.onego.ru/smalt/dtd.

Заключение. Создан формат описания SMR.DTD v1.0 (версия 1.0), который является языком на основе XML, предназначенным для проведения и хранения синтаксического и морфологического разбора произведений, а также представления полученных разборов в том виде, который необходим для текущей работы исследователя. Формат выложен в Интернет для публичного доступа и обсуждения.

В настоящее время исследуется возможность трансформации полученных разборов в любую необходимую форму с использованием XSLT-преобразований (Extensible Stylesheet Language Transformations). Это требуется для выделения только требуемого подмножества данных, например, только синтаксического или морфологического разбора.

СПИСОК ЛИТЕРАТУРЫ

1. Захаров В. Н., Леонтьев А. А. , Рогов А. А. , Сидоров Ю. В. Программная система поддержки атрибуции текстов статей Ф. М. Достоевского // Труды Петрозаводского государственного университета. Сер. Прикладная математика и информатика. Петрозаводск, 2000. Вып. 9. С. 67–80.
2. Розенталь Д. Э., Голуб И. Б., Теленкова М. А. Современный русский язык. — М. : Айрис-пресс, 2002. — 448 с.
3. Рогов А. А., Сидоров Ю. В., Король А. В. "СМАЛТ" — от построения корпуса текстов к способам их обработки статистическими и эвристическими методами // Региональная информатика-2004 "РИ-2004": материалы IX Санкт-Петербургской международной конференции, Санкт-Петербург, 22–24 июня 2004. СПб, 2004. С. 243–244.
4. Rusty H. E., Scott M. S XML in a Nutshell — USA : O'Reilly, 2003. — 300 p.

Суровцова Т. Г. — МКО — 2007, т. 2, стр. 254–261
Surovtsova T. G. — MCE — 2007, v. 2, p. 254–261

THE DESCRIPTION OF SYNTACTIC AND MORPHOLOGICAL ANALYSIS OF THE TEXT WITH USE XML

Surovtsova T. G.

(Russia, Pushchino)

The format for the description of syntactic and morphological analyses of texts according to grammar of Russian on XML is offered. DTD-format description is created, and it is possible to spend analysis of the text. This format is offered for public use.