Универсальная система разметки текста ATE-2

А. И. Зобнин, А. В. Маркелова
Институт русского языка им. В. В. Виноградова РАН, МГУ им. М. В. Ломоносова, Москва, Россия

We present a new version of the Ancient Text Editor (ATE-2) which is to replace the previous version of the syntactical editor used in Vinogradov Institute for Russian Language of RAS. It relies on the object-oriented concepts and allows the user creation of his own markup templates. These templates define the rules of creating, managing, sorting, grouping and viewing the objects. The system can transform data from one set of templates into another. The program is based on the Microsoft .NET technology and relational database management systems.

Система синтаксической разметки древнерусских текстов, существовавшая в Институте русского языка им. В. В. Виноградова РАН на 2005 г., требовала существенной модернизации. Эта система позволяла сопоставлять определенному связному фрагменту текста набор признаков из заданного списка, а также строить запросы [Зобнин 2005: 44–47]. Несмотря на то, что с ее помощью можно было получить интересные результаты, она обладала рядом недостатков:

– нельзя было объединить в единый фрагмент отдельно расположенные части текста;

– невозможно было указать роль отдельных частей фрагмента во всем фрагменте; система не могла устанавливать связи между выделенными фрагментами.

У пользователей системы не было четкого согласия относительно границ выделяемых фрагментов. В итоге в разных частях текста однотипные структурные синтаксические элементы то включались, то не включались во фрагмент (например, предлог при существительном во фрагменте ‘*субстантив — атрибут*’).

Система была написана на Microsoft Access и имела не очень удобный интерфейс.

Было замечено, что особенностью синтаксиса древнерусского языка является существование «нелинейных» синтаксических отношений. Использование классического синтаксического дерева значительно искажает реальные синтаксические отношения в тексте, поэтому требовалось реализовать сетевую структуру данных.

На разработчиков системы большое влияние оказала система «Манускрипт», созданная группой разработчиков из ИжГТУ и УдГУ под руководством проф. В. А. Баранова [Баранов 2003: 159–165; Баранов 2003: 234–270]. В Институте русского языка была предпринята попытка частично реализовать модель синтаксической разметки этой системы. Однако, при всей универсальности предложенной модели, эта реализация оказалась слишком громоздкой. Возникало много вопросов о том, как правильно надо размечать данный фрагмент, что свидетельствовало о неоднозначной интерпретации модели и некоторой несогласованности между разработчиками.

В ходе экспериментальной работы было принято решение о том, что указание компонентов связей и их свойств должно предшествовать указанию свойств самих связей и синтаксических единиц. Кроме того, возникла идея использовать только бинарные связи, а запросы к системе строить как условия на комбинацию таких бинарных связей. Но последний подход является противоположной крайностью, которая обладает теми же недостатками: система не является гибкой, она громоздка, рано или поздно возникнет вопрос о совместимости с другими системами, требуется создание отдельного модуля запросов.

Было найдено решение, устраняющее все перечисленные недостатки и сохраняющее все достоинства. На основе новой предлагаемой модели разметки можно реализовать (пока теоретически) все перечисленные выше модели. Кроме того, в рамках новой модели возможно сочетание синтаксической, морфологической и прочих разметок текста. Сформулируем основные возможности, реализованные в новой системе синтаксической разметки ATE-2. Поскольку используемая нами терминология еще не устоялась, то в скобках мы будем указывать альтернативные названия, пришедшие из объектно-ориентированного программирования.

Система построена на основе «гибких» макетов. Пользователь имеет возможность самостоятельно задавать структуру разметки,

определять шаблоны (за исключением встроенных шаблонов) и их компоненты, конструировать запросы. При загрузке размеченного материала сначала загружается макет разметки, и на его основе уже отображается и ведется сама разметка.

Система использует *объектно-ориентированный подход*. *Макет* (структура разметки) представляет собой набор *шаблонов* (классов) — некоторых абстрактных типов данных, которые обладают определенными *компонентами* (членами класса, полями, элементами шаблона) или *наборами* таких компонент. Каждая компонента шаблона — это снова некоторый шаблон (или встроенный тип данных — текстовая строка, число и т. д.). Все шаблоны имеют обязательную текстовую компоненту '*содержание*' (она может вычисляться по определенным правилам через содержание компонентов), а также могут иметь поля, предназначенные для сортировки и сравнения. Процесс разметки состоит в создании *объектов* шаблонов (экземпляров, единиц) с указанием уже существующих единиц в качестве элементов данного объекта шаблона. Кроме информации о типах своих компонентов, в шаблоне могут присутствовать *ограничения* на их значения. Ограничения имеют универсальный характер и могут задаваться в виде логических выражений первого порядка, которые должны быть истинными для любого объекта шаблона. Можно провести следующую аналогию с реляционными базами данных: шаблоны играют роль таблиц и связей между ними, элементы шаблонов аналогичны полям таблицы (при этом возникают связи «один ко многим» и «много ко многим»). Ограничения напоминают условия целостности базы данных. Однако возможности объектно-ориентированного представления данных с указанием ограничений намного удобнее и шире, чем при реляционном подходе. Еще одним существенным отличием является возможность *наследования* одних шаблонов из других (то есть шаблон-наследник приобретает все компоненты и ограничения шаблона-предка, при этом, возможно, добавляя что-то свое). Таким образом, макет разметки — это определенная иерархия шаблонов. Кроме того, можно задать правила сортировки и группировки элементов в наборе компонент.

В системе нет структурных различий между элементами разметки и запросами к базе данных. Любой запрос может быть представлен в виде шаблона, компонентами которого являются интересующие пользователя элементы разметки, а ограничения задают критерии отбора данных. Таким образом, запрос — это всегда выборка экземпляров определенного шаблона. Однако, в отличие от процесса разметки, элементы, выбираемые запросом, не обязаны заранее физически храниться в базе данных — они могут динамически создаваться на время просмотра запроса.

Ограничения, задаваемые в шаблонах, должны быть предназначены не только для поддержки корректности, но и для автоматизации процесса разметки. Так, пользователь может лишь указать некоторый набор объектов (например, выделив их мышью), а программа должна предложить ему список тех шаблонов, обязательными компонентами которых эти объекты могут являться (с соблюдением ограничений на свойства этих объектов). Пользователь выбирает подходящий шаблон из списка и создает новый объект. Затем он может выбрать значения дополнительных свойств (атрибутов) этого объекта. Список допустимых значений каждого необязательного свойства может быть получен через ограничения и уже заданные свойства. Если окажется, что допустимое значение данного свойства единственно, то программа выбирает это значение автоматически.

Система имеет возможность преобразования одних макетов в другие. Тем самым исчезает вопрос о совместимости форматов. Каждая группа пользователей ведет работу на основе своего макета, а затем, при необходимости преобразовать данные в иной формат, пользователем указываются «правила перехода» от одного макета к другому, то есть правила преобразования шаблонов одного макета в шаблоны другого.

Система имеет удобный интерфейс. В результате время, затрачиваемое пользователем на разметку текста, существенно сокращается.

Шаблон напоминает многополярную связь из системы «Манускрипт», однако, в отличие от связи, он имеет собственный набор свойств и обладает ограничениями. С другой стороны, синтаксическую разметку системы «Манускрипт» можно смоделировать в рамках нового подхода. Можно создать макеты и для модели бинарных связей, и для предыдущей модели разметки синтаксиса ATE.

Очередная версия новой системы ATE-2 (Ancient Text Editor -2) реализована на платформе Microsoft .NET с использованием реляционных баз данных Microsoft Access (для автономной работы) и Microsoft SQL Server.

Список литературы

Баранов и др. 2003а — *Баранов В. А.* Специализированной текстовый редактор «Манускрипт» Системы обработки древних рукописей / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, А. Н. Миронов, В. А. Романенко // Информационный бюллетень Ассоциации «История и компьютер». — 2003. — № 31. — С. 159–165. (http://manuscripts.ru/mns/docs/AIK2003b.pdf).

Баранов и др. 2003b — *Баранов, В. А.* Электронные издания древних письменных памятников и технология создания полнотекстовых баз данных / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, О. В. Зуга, А. Н. Миронов [и др.] // Круг идей: электронные ресурсы исторической информатики : тр. VIII конф. Ассоциации «История и компьютер» / под ред. Л. И. Бородкина, В. Н. Владимирова. — М. ; Барнаул : Изд-во Алт. ун-та, 2003. — С. 234–270. (http://manuscripts.ru/mns/docs/AIK2003.pdf).

Зобнин 2005 — *Зобнин, А. И.* Корпус древнерусских переводов XI–XII веков: результаты и перспективы / А. И. Зобнин, А. А. Пичхадзе // Научно-техническая информация. Информационные процессы и системы. — 2005. — Сер.2. — № 3. — С. 44–47.

Кагарлицкий и др. 2003 — *Кагарлицкий, Ю. В.* Проблемы создания электронного корпуса переводных памятников древнерусской письменности XI–XII вв. / Ю. В. Кагарлицкий, А. А. Пичхадзе, С. А. Шаров // Корпусная лингвистика в России. — 2003. — Вып. 10. — Сер. 2.

Интернет-издание «Пантелеймоново Евангелие XII–XIII вв. (РНБ, Соф. 1)»

О. В. Зуга
Удмуртский государственный университет, Ижевск, Россия

The article contains the description of the Internet site «The Panteleymon Gospel (RNB, Sof. 1)» which presents the full text of the manuscript (transcription), help materials, textual both linguistic researches and comments. The basis of the edition — the text-through database — includes some kinds of the full text, the grammatical information on the linguistic units and comments and gives the user the opportunity of preparation of queries, fast search of the data and its sorting.

В XIX веке публикацией древних богослужебных рукописей занимались многие крупнейшие представители русской лингвистической науки — А. Х. Востоков, И. В. Ягич и др. [Востоков 1843; Ягич 1886]. В конце XX — начале XXI века вновь стали появляться публикации и исследования евангелий, миней, триодей и других церковных текстов (см., например, [Жуковская 1997; Алексеев 1981, 1988; Нечунаева 1994; Баранов 2003; Крысько 2005]).

Однако до сих пор большая часть уникальных рукописных памятников, значительных по объему, разнообразных по языку и стилистическим ресурсам, не издана. Нехватка новых источников, доступных в обработанном, то есть опубликованном виде, и в то же время обилие письменных славянских рукописей, «разбросанных» по различным хранилищам, ставят серьезные препятствия на пути овладения письменным богатством во всей его полноте и влекут за собой, с одной стороны, наблюдаемое в последние годы снижение интереса к вопросам исторической грамматики церковнославянского и русского языков, а с другой — появление многочисленных, пестрых и взаимоисключающих концепций относительно природы литературно-письменного языка у восточных славян в эпоху средневековья (см. об этом [Алексеев 1988: 28–30]). Кроме того, в настоящее время уже далеко не все находящие-