Универсальная система синтаксической разметки текста ObjectATE

А. И. Зобнин (Механико-математический ф-т МГУ им. М. В. Ломоносова, Alexey.Zobnin@gmail.com)
А. В. Сахарова (Институт русского языка им. В. В. Виноградова РАН, nenen@mail.ru)

Система ObjectATE

- разрабатывается и используется уже более двух лет в Отделе лингвистического источниковедения института русского языка им. В. В. Виноградова РАН для морфологической и синтаксической разметки древнерусских текстов: переводных памятников и летописей;
- является гибким и многофункциональным средством создания лингвистических текстовых корпусов;
- позволяет заниматься лингвистической разметкой текста начиная с того уровня, который выбирает пользователь, и по тем параметрам, которые он задает сам;
- использует объектно-ориентированный подход к представлению данных;
- позволяет получать выборки по запросам к данным, а также наглядно визуализировать разметку с помощью деревьев, указателей и т. п.

Разметка текста

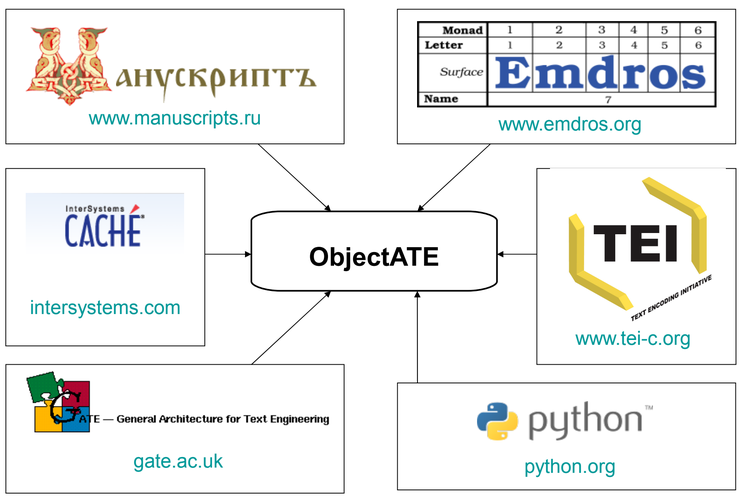
Разметка древнерусских текстов в системе ObjectATE ведется в ручном режиме по следующим причинам:

- корпус невелик;
- язык изучен недостаточно глубоко (к примеру, не существует словарей управлений);
- тексты отличаются друг от друга по орфографии, морфологии и по многим синтаксическим параметрам;
- тексты не свободны от разного рода нарушений грамматической связности, темных мест, описок.

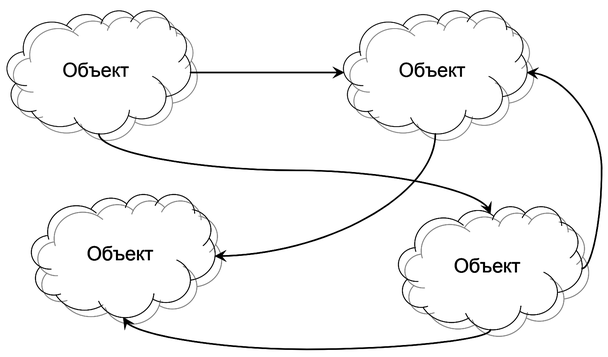
Однако модель системы позволяет накладывать жесткие условия на разметку для контроля корректности и выбора подходящих вариантов.

Эта модель не противоречит возможной автоматизации разметки в будущем.

Системы, оказавшие влияние на разработку



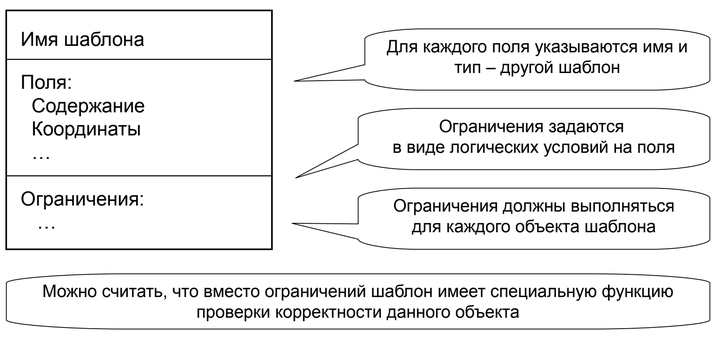
Объектная модель данных



Документ является набором связанных друг с другом объектов разметки. Разметка состоит в создании новых и изменении существующих объектов.

Объекты и шаблоны

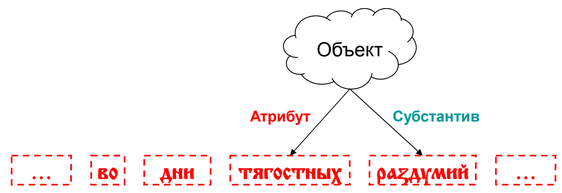
Шаблон (класс, тип) в ObjectATE – аналог класса в программировании. Он определяет общие черты и поведение объектов, а также содержит *ограничения* — логические условия (инварианты).
Объект – конкретный экземпляр (элемент разметки), созданный по шаблону.



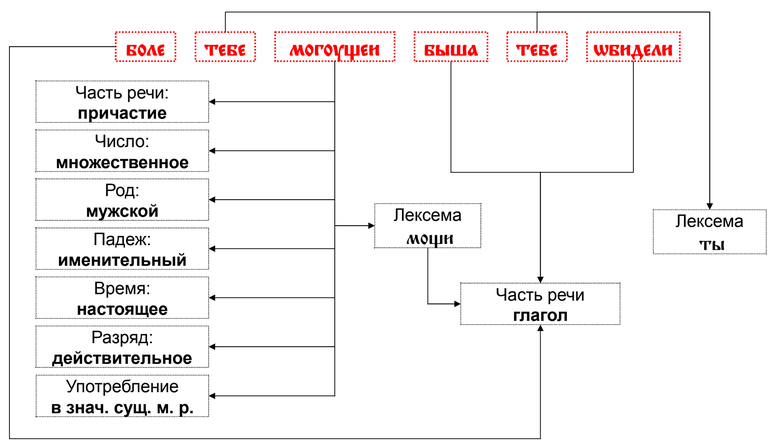
Объекты можно мыслить как точки или отрезки на прямой, которые можно сравнивать по их координатам.

Пример шаблона

Связь с согласованным атрибутом
Содержание Координаты Атрибут: Словоформа Субстантив: Словоформа
Атрибут.[Часть речи] IN {'прилагательное', 'местоимение', 'числительное'} Субстантив.[Часть речи] = ‘существительное’ Атрибут.Падеж = Субстантив.Падеж



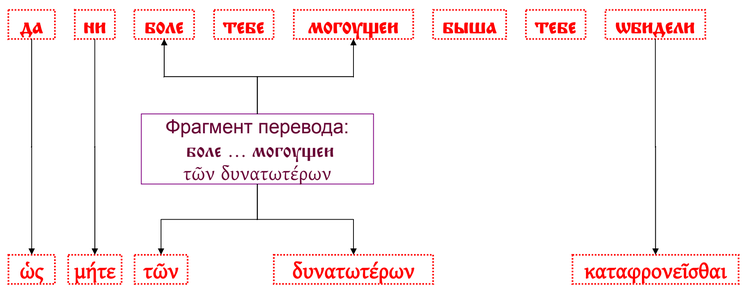
Пример: морфология



Пример: синтаксис

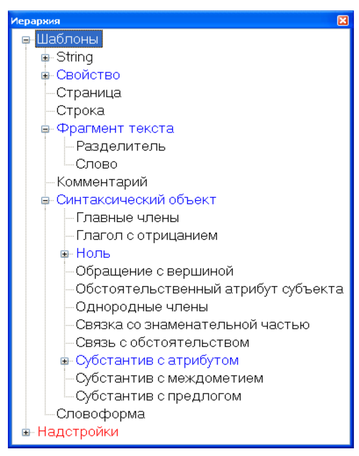
Стрелка обозначает связь между объектами (указывает на значение поля).

Пример: перевод



Сопоставление текстов используется для сравнения порядка слов в синтаксических конструкциях оригинала и перевода.

Метаданные



Метаданные – иерархия шаблонов, которые используются для разметки.

Шаблоны выстраиваются в иерархии множественного наследования. Наследник приобретает все поля и ограничения предков.

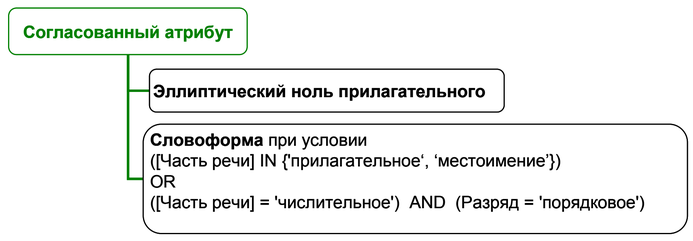
Абстрактные шаблоны нужны для описания узлов иерархии; они не могут порождать новые объекты.

В системе ObjectATE метаданные могут разделяться различными источниками данных.

Надстройки

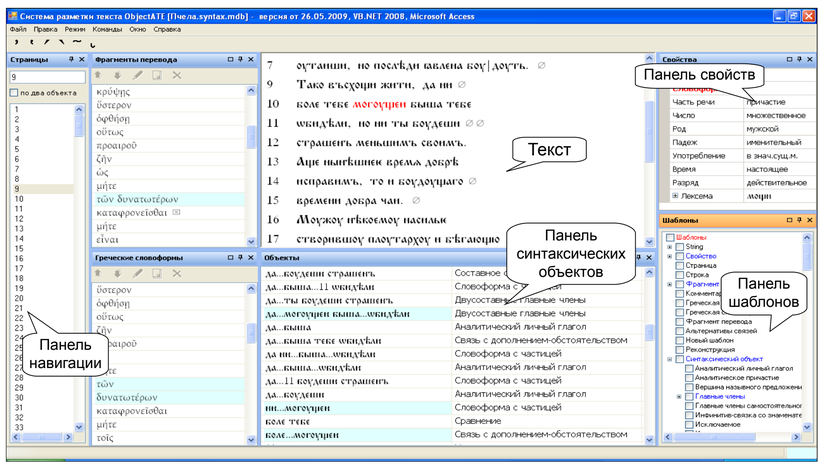
Надстройка над шаблоном – категория, к которой могут относиться объекты шаблона при выполнении определенных условий.

Надстройки нужны как для описания запросов, так и для уточнения ограничений шаблонов.



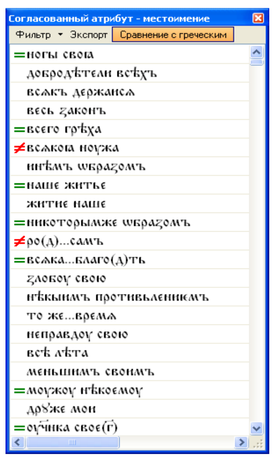
В этом примере к надстройке «согласованный атрибут» относятся любые эллиптические нули прилагательного и те словоформы, для которых выполнено условие.

Редактор ObjectATE



Редактор написан в среде Microsoft .NET Framework. Он имеет удобный интерфейс с «плавающими» окнами. Содержимое окон, и связи между ними описываются внешним образом в конфигурационных файлах.

Выборки



Результаты запроса на все объекты типа «связь с согласованным атрибутом» с условием

```
Атрибут.[Часть речи]='местоимение'.
```

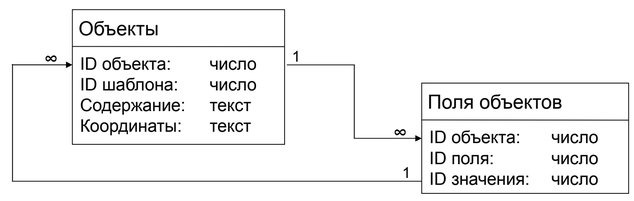
В окне с результатами также показано сравнение с порядком слов в греческом тексте (отмечены совпадения и несовпадения в случаях, когда сравнение возможно).

Система ObjectATE позволяет фильтровать выборки по текущей области просмотра, экспортировать результаты в текстовый файл и представлять их в иерархическом виде.

Хранение данных

Объектная модель данных абстрагируется от конкретного способа их хранения и обработки. В редакторе ObjectATE для хранения объектов используется реляционная база данных, а метаданные и конфигурационные настройки хранятся как в базе данных, так и в XML-файлах.

Схема реляционной базы данных:



Логические условия и ограничения переводятся системой на язык запросов SQL.

Результаты

Разработана объектная модель данных, подходящая для любых уровней лингвистической разметки текста по тем правилам, которые определяет сам пользователь.

Создан специальный редактор ObjectATE, позволяющий вести такую разметку и наглядно ее визуализировать, настраивать метаданные, получать выборки результатов запросов.

Система ObjectATE позволила на совершенно новом уровне заниматься разнообразными лингвистическими и текстологическими исследованиями анализируемых текстов, формулируя разнообразные запросы к синтаксически размеченному корпусу.

По ней уже изучается стратегия работы древних переводчиков, например закономерности изменения ими порядка слов оригинального текста.