Возможности применения универсальной системы синтаксической разметки текста ObjectATE

А. В. Сахарова

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Система разметки текста ObjectATE используется в Отделе лингвистического источниковедения ИРЯ РАН для создания и хранения синтаксического описания древнерусских памятников: в данный момент – переводной антологии «Пчелы» и Киевской летописи по Ипатьевскому списку.

Наша система является очень гибкой и многофункциональной. Предназначена она, прежде всего, для ручной разметки текста, что в первую очередь и требуется в случае, когда корпус текстов невелик и не свободен от разного рода описок, двусмысленных и вообще темных мест и язык, на котором он написан, является мертвым. Ее главной отличительной чертой является возможность самому пользователю описывать правила разметки в рамках общей модели. В системе можно просто в полностью ручном режиме размечать морфологию для предварительно уже разделенного на словоформы текста: присваивать словоформам значения морфологических категорий. Лемматизацию, автоподстановку морфологических параметров, создание словников и указателей предполагается внедрить в эту систему в ближайшее время, а сейчас для этих целей в Отделе лингвистического источниковедения применяется другое программное средство – Редактор древнерусских текстов ATE. Также разрабатывается способ отмечать разного рода неправильности и темные места текста, связанные с переосмыслениями писцов, разного рода анаколуфами и т. п. Даже при отсутствии морфологической информации (т.е. если текст всего-навсего разделен на словоформы) можно сразу заниматься ручной синтаксической разметкой текста, т.е. созданием в базе данных новых объектов – единиц синтаксического анализа, или синтаксических объектов.

Синтаксический анализ, как известно, может проводиться по группам и по зависимостям. Для того чтобы разбирать текст по зависимостям, можно просто создать список всех необходимых типов синтаксических связей и

начать создавать («склеивать») синтаксические объекты, состоящие из пары словоформ – вершинной и подчиненной ей.

Можно заниматься анализом по группам, собирая точно такие же синтаксические объекты, но не только из словоформ, но и из самих синтаксических объектов. Для этого уже необходимо применение механизма так называемых надстроек. Надстройка – это описываемое пользователем множество словоформ и синтаксических объектов, обладающих определенными свойствами. Например, надстройка «Сказуемое» должна включать в себя и одну словоформу, и синтаксический объект, состоящий из глагола *быти* и знаменательной части, чтобы и то, и другое, допустим, можно было соединить с подлежащим для создания нового объекта – связи «Подлежащее–сказуемое».

Можно, впрочем, и комбинировать оба метода синтаксического анализа, как это делаем мы: некоторые синтаксические группы собираем целиком, но для большинства случаев ограничиваемся построением связи между вершинами.

Однако если морфологическая информация о словоформах для разбираемого текста присутствует (так, для разбираемых нами «Пчелы» и Киевской летописи морфология уже заранее сделана в другой системе), то при синтаксической разметке (что по зависимостям, что по группам) ее можно использовать, тем самым эту разметку упрощая или, по крайней мере, снижая вероятность ошибки.

Для этого нужно также использовать механизм надстроек, так как он позволяет задавать условия на морфологические свойства словоформы, позволяющие ей играть определенную синтаксическую роль. Предположим, для словоформы, входящей в надстройку *Сказуемое*, можно оговорить, что она должна представлять собой личный глагол. Можно также задавать условия на вхождения целой синтаксической группы в надстройку (как правило, это бывают условия на ее вершину). Так, можно сформулировать, что для составного сказуемого вершиной должен быть глагол *быти* в личной форме, хотя возможны условия и не только на вершину. Для удобства пользователей при описании подобных условий предполагается ввести для

синтаксического объекта понятие вычисляемых полей, т.е. механизма, выводящего значения параметров самого синтаксического объекта из определенных параметров его полей.

Описывая условия на вхождение словоформ и синтаксических объектов в определенное множество, мы можем использовать это множество не только для создания новых синтаксических объектов, но и для вызова из базы данных списка всех объектов, удовлетворяющих сформулированным нами условиям (например, создать надстройку «Составное сказуемое с нулевой связкой» и сразу же взглянуть на все подобные сказуемые).

Также можно задать не только условия вхождения словоформы или синтаксического объекта в надстройку, но и ограничения самого синтаксического объекта – условия сочетания свойств в него входящих словоформ. Скажем, для подлежащего и сказуемого оговаривается согласование по лицу в следующем виде: если лицо сказуемого первое, то подлежащее либо тоже стоит в первом лице, либо представляет собой синтаксический ноль, и т. д.

Все это означает, что так как мы располагаем морфологической разметкой и системой описания простейших синтаксических правил языка (основных правил согласования и управления, связанных с морфологическим обликом словоформ, а также с порядком слов), то уже гораздо меньше вероятность допустить ошибку при анализе. Синтаксический объект не создастся, если его части не войдут в соответствующие надстройки (скажем, существительное в косвенном падеже не будет трактоваться системой как подлежащее) или если эти части войдут, но ограничения самого объекта не позволят создать синтаксический объект (если, например, имя стоит в именительном падеже, но не согласовано с личным глаголом по лицу, из них не получится создать объект «Связь подлежащее–сказуемое»).

В будущем для системы планируется создание механизма пересчета анализа по зависимостям в анализ по группам, т.е. создание механизма сбора группы по вершине (например, всего простого предложения по его сказуемому).

Накопленная информация о простейших синтаксических правилах языка может позволить в дальнейшем сделать синтаксический сбор полуавтоматическим, т.е., используя конструктор объектов, сразу размечать определенные фрагменты текста по сформулированным правилам.

Наконец, так как некоторые из текстов, с которыми мы имеем дело, представляют собой переводы, система ObjectATE предусматривает и средства описания соответствий между оригиналом и переводом, т.е. исследования характера переводческой деятельности создателей памятника. База данных по «Пчеле» содержит параллельный греческий текст и список греческих лексем, что позволяет разметчику создавать особые объекты лингвистического анализа, так называемые фрагменты перевода, ставя в соответствие одной или нескольким славянским словоформам одну или несколько словоформ оригинала, а также, если нужно, лексемы оригинала.

Список литературы

Зобнин и др. 2006 – *Зобнин, А. И.* Универсальная система разметки текста АТЕ-2 / А. И. Зобнин, А. В. Маркелова // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы междунар. научн. конф., Ижевск, 13-17 июля 2006 г. – Ижевск, 2006. – С. 51–55.

Possibilities for applying ObjectATE, a universal system for syntactic markup of text

Anna V. Sakharova

Vinogradov Institute of the Russian Language of the Russian Academy of Sciences, Moscow, Russia

The linguistic opportunities provided by Object ATE, a universal syntactic annotation system, are described. The system allows the user to perform morphological and syntactic analysis of ancient texts. In the future a semi-automatic parser of Old Russian may be created. In cases where the text analyzed is a translation, the system will describe the degree of conformity of the translation to the original.