Разметка корпуса Азбуковника 1596 г. на основе рекомендаций TEI

Digitalising of the Russian Manuscript Dictionary of 1596 Using the TEI Encoding Scheme

Коваленко К.И., магистр, мнс
kira.kovalenko@gmail.com
Институт лингвистических исследований
Российской академии наук,
Санкт-Петербург, Россия

Ключевые слова: историческая лексикография, азбуковники, разметка текстов, TEI

The paper deals with the problems of Russian manuscript dictionary digitalising. The dictionary has a complicated entry structure, which combines different types of information: word origin, definitions, quotations from literary texts, references to other words, etc. It is supposed to use the TEI encoding scheme, as it provide a wide variety of tags for manuscripts and dictionaries encoding. The digitalising would provide more effective access to the information and accelerate linguistic researches.

Создание корпусов текстов — одно из наиболее перспективных направлений современной лингвистики. В настоящее время в нашей стране активно развивается проект Национального корпуса русского языка, одно из направлений которого предполагает расширение корпуса за счет включения текстов, отражающих историю русского языка, — от берестяных грамот и летописей до текстов XVIII — начала XIX вв. В частности, предполагается создать Корпус старорусских текстов XV-XVII вв., в рамках которого ожидается пополнение корпуса непосредственно за счет текстов в электронном виде, а также разработка компьютерной морфологии и словаря старорусских текстов [http://www.corpling-ran.ru/n2.html]. Можно предположить, что корпус среднерусских текстов стал бы более репрезентативным, если бы в него были вклю-

чены тексты азбуковников XVI-XVII вв. — предшественников современных словарей.

Азбуковник как лексикографический жанр сложился к середине XVI в. на основе более ранних словарных сводов — ономастиконов (глоссариев имен библейских персонажей и наименований библейских мест), приточников (перечней символов Псалтири), глоссариев к «Лествице» Иоанна Синайского и словарей-разговорников [История русской лексикографии 2001: 40-49]. В дальнейшем азбуковники активно пополнялись новыми статьями, источником которых служили глоссы на полях рукописей, пояснения непонятных читателю реалий в текстах литературных произведений. Поэтапный характер формирования азбуковников, разнородные сведения, представленные в нем, являются причиной того, что статья азбуковника включает в себя достаточно разнородную информацию и далеко не всегда имеет четко выстроенную линейную структуру.

Азбуковник, выбранный для разметки, является типичным представителем своего жанра. Он был создан в Новгороде в 1596 году в монастыре Антония Римлянина и сохранился до настоящего времени в единственном списке начала XVII в. (РНБ, собр. Погодина, № 1642). В нем содержится более 6000 статей, которые организованы по первой букве и по следующей гласной. В отличие от своих предшественников, в нем представлена не только лексика греческого, латинского, старославянского и тюркского происхождения, но имеется значительный пласт лексических единиц из польского языка.

Структура статьи азбуковника имеет достаточно сложную организацию. Так, в заголовочной части статьи может быть слово, словосочетание или целая фраза (например: *Кеси е҆ро́тисо^{н} а҆паоу҆тонъ* или *Что е^{с} пмть чювьствъ дше҃вны^{х} и҆ пмть тѣле́сныхъ*). Зона толкования может содержать объяснение сразу нескольких слов (*Житїе и жизнь (т) житїе нарицае^{т} е҆же кто какова ѿца҆ и҆ коее҆гоҀ града и҆ коее҆мҀ вѣ́ры . жи́знь же се е^{с} каковыми дѣлы бг҃у оу҆годиҀ, Ї коими́ дарова́нїи ѿ га҃ прославле^{н}, и҆ ка́ко теченїе подвига скончаҀ*), указание на паронимы (*Гранограф^{ѳ} (т) гране́и писе^{ц} землемѣ́рны^{х}. їно бо е^{с} гранограф^{ѳ}. ї и́но хронограф^{ѳ}*), а также дополнительную информацию, относящуюся к толкованию данного слова опосредованно (*Всеплóдїе (т) в҆ветхо^{м} законѣ прїведе́ное*

о҆вча. и҆ли волъ на жр ̑тву вг ̑ви. е҆гда̀ закла́вше всесо́жгу ᵐ, то̑ всепло́дїе ї́менуе^т^сѧ. а҆ е҆гда̀ заклавше, нѣ́кїѧ ча́стї бг̑у ѿдѣла^т^ нѣкїѧ же ї҆е҆реѡ^м^ ї народо^м^ на снѣденїе, то нариче^т^сѧ же́ртва). Как правило, над заготовочным словом указывается предполагаемый язык-источник, однако данная помета может присутствовать и над любым другим словом в любой части статьи (*Салафѳа {ж}, и҆ри́на {г}, ҆е́же е^с^ мирнаѧ*). В некоторых случаях на полях указываются литературный источник (источники), из которых данная статья попала в азбуковник. Но иногда ссылка на источник оказывается без пометы, и остается только догадываться, к какой же статье азбуковника она относится.

Разметку корпуса словаря предполагается осуществить на основе рекомендаций TEI — Text Encoding Initiative (версия P5). На базе рекомендаций TEI уже были реализованы и продолжают функционировать более 150 проектов [http://www.tei-c.org/Activities/Projects]. Среди отечественных проектов можно отметить информационно-поисковую систему "Русская литература XVIII века" [http://antology-xviii.spb.ru], также выдвинут ряд предложений по использованию TEI в работах [Вадяев 2005; Вотинцев 2006; Бабалык, Варфоломеев, Пигин 2010; Захаров, Митрофанова, Михайлова 2011].

Все более широкое распространение стандартов TEI обусловлено тем, что рекомендации по разметке рассчитаны на разные типы текстов: прозаические и стихотворные произведения, словари, транскрипцию разговорной речи. Отдельное внимание уделяется электронному представлению текстов рукописных источников. Необходимо также отметить, что рекомендуемая разметка рассчитана в том числе и на словари со сложной структурной организацией, что дает возможность пользоваться исключительно предложенными тегами и атрибутами, не усложняя код разметки дополнительными нововведениями.

В качестве примера использования предлагаемых TEI тегов возьмем статью азбуковника *Кида́ръ {ж} {ѱлт̑о^м^ рѳї} (т) тма. и҆лѝ клобукъ а҆рхїе҆ре́искїи. ҆и́же и҆ митра гл̑е/тсѧ. Ї па́ки кида^р^ нарица́е^т^ и҆ до́мы татар̑скїѧ. о҆вы бо ҆и́хъ в землѝ и҆ско́паны , и҆ того рад“ те́м̑ныи; о҆вы н а колесни́ца ^х^, по^п^стмѝ о҆граждє́ны; ’е́сть же и҆ гра ^д^ нарица́емъ кида́ръ о҆ нем̑же пи́шетъ , всели́хсѧ с с елы кидаръскї́ми.* Разметка данной статьи, ориентированная на структурную организацию, будет выглядеть следующим образом:

<entryFree>

<form><w>Кидáръ<note><lang>ж</lang><bibl><title>ѱлѣ҃ом рѳї҃</title></bibl></note></w></form>

<sense> (т) <def n="1">тма.</def> и҆ли <def n="2"> клобукъ а҆рхїе҆рéискїи. ꙗ҆же и҆ <mentioned>митра</mentioned> гл҃етсѧ.</def> Ї пáки кидар нарицáет и҆ <def n="3">дóмы татар҆скїѧ. о҆вы бо ꙗ҆хъ в землѝ и҆скóпаны, и҆ того рад“ тéмныи; о҆вы на колесни́цах, полстмѝ о҆гражде́ны;</def> <def n="4">ꙗ҆́сть же и҆ град нарицáемъ кидáръ о҆ немѓже пи́шетъ, <cit> <quote>всели́хсѧ с селы кидаръскїми</quote> </cit></def></sense>

</entryFree>

Кроме разметки текста в соответствии со структурой статьи, предполагается ввести дополнительные сведения: по возможности определить иноязычное слово в языке-источнике, лексикографический или литературный источник и контекст, послуживший материалом для создания той или иной статьи, тематическая разбивка лексики по группам. Это должно значительно облегчить лингвистическое исследование корпуса азбуковника, в частности, поможет точнее определить, из каких языков лексика той или иной тематической группы наиболее активно проникала в русский язык, а также выявить фонетические и графические приемы передачи иноязычных слов.

Литература

Бабалык, Варфоломеев, Пигин 2010 — Бабалык М.Г., Варфоломеев А.Г., Пигин А.В. Использование формата TEI для публикации и анализа списков произведений вопросо-ответного жанра // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28-31 октября 2010 г.) / отв. ред. В.А. Баранов. Уфа; Ижевск: Вагант, 2010. С. 17-20.

Вадяев С.Е. Лингвистические принципы построения и использования корпуса текстов для исследования официально-делового стиля современного немецкого языка (на материале электронного корпуса «DER»). Автореф. дис. на соиск. учен. степ. к.филол.н. Н. Новгород, 2005.

Вотинцев П.А. Использование формата TEI для обмена данными с полнотекстовой информационно-поисковой системой «Манускрипт» // Современ-

ные информационные технологии и письменное наследие: от древних рукописей к электронным текстам: материалы междунар. науч. конф. (Ижевск, 13-17 июля 2006 г.) / отв. ред. В.А.Баранов. Ижевск: Изд-во ИжГТУ, 2006. С. 30-31.

Захаров, Митрофанова, Михайлова 2011 — Захаров В.П., Митрофанова О.А., Михайлова В.Д. Разметка словарей в соответствии со стандартом TEI // Информационные технологии в лексикографии. СПб.: СПбГУ, Филологический факультет, 2011. С. 61-69.

История русской лексикографии 2001 — История русской лексикографии / отв. ред. Ф.П. Сороколетов. СПб.: Наука, 2001.