МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

Edited with the demo version of
Infix Pro PDF Editor
To remove this notice, visit:
www.iceni.com/unlock.htm

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12–06–06061–г),
гранта РГНФ (проект № 12–04–14154–г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:
В. А. Баранов, д-р филол. наук, проф.
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

И741 **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

© Петрозаводский государственный университет, 2012
© Ижевский государственный технический университет им. М. Т. Калашникова, 2012

Edited with the demo version of Infix Pro PDF Editor
To remove this notice, visit: www.iceni.com/unlock.htm

РАЗМЕТКА КОРПУСА АЗБУКОВНИКА 1596 ГОДА НА ОСНОВЕ РЕКОМЕНДАЦИЙ TEI

К. И. Коваленко

Институт лингвистических исследований Российской академии наук, Санкт-Петербург

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 digital-ising 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 направ-лений которого предполагает расширение корпуса за счет включения тек-стов, отражающих историю русского языка, — от берестяных грамот и лето-писей 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 тек-стов (http://www.corpling-ran.ru/n2.html). Можно предположить, что корпус среднерусских текстов стал бы более репрезентативным, если бы в него были включены тексты азбуковников XVI–XVII вв. — предшественников со-временных 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), приточни-ков (перечней символов Псалтири), глоссариев к «Лествице» Иоанна Синай-ского и словарей-разговорников [История, 2001: 40–49]. В дальнейшем азбу-ковники активно пополнялись новыми статьями, источником которых слу-жили глоссы на полях рукописей, пояснения непонятных читателю реалий в текстах литературных произведений. Поэтапный характер формирования аз-буковников, разнородные сведения, представленные в нем, являются причи-ной того, что статья азбуковника включает в себя достаточно разнородную информацию и далеко не всегда имеет четко выстроенную линейную струк-туру.

Азбуковник, выбранный для разметки, является типичным представителем своего жанра. Он был создан в Новгороде в 1596 г. в монастыре Антония Римлянина и сохранился до настоящего времени в единственном списке на-чала 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.

Edited with the demo version of Infix Pro PDF Editor
To remove this notice, visit: www.iceni.com/unlock.htm

Структура статьи азбуковника имеет достаточно сложную организацию. Так, в заголовочной части статьи может быть слово, словосочетание или целая фраза (например: *Кеси ерописо н апаоутонъ* или *Что е с пꙗть чювьствъ дше҃вны х и пꙗть тѣлесныхъ*). Зона толкования может содержать объяснение сразу нескольких слов (*Житїе и жизнь (т) житїе нарицае т еже кто какова ѿц҃а и коего града и коеꙗ в ѣры. жизнь же се е с каковыми дѣлы бг҃у оугоди , Ї коими дарованїи ѿ га҃ прославле н, и како теченїе подвига сконча*), указание на паронимы (*Гранограѳ (т) гранеи писе ч землемѣрны х. їно бо е с гранограѳ. ї ино хронограѳ*), а также дополнительную информацию, относящуюся к толкованию данного слова опосредованно (*Всеплодїе (т) в ветхо м законѣ прїведеное о вча. или волъ на жр҃тву вг҃ви. егда заклаше всесожгу т, то всеплодїе їменуе тсꙗ. а егда заклавше, н ѣкї ꙗ часті бг҃у ѿдѣла т нѣкїꙗ же їерew м ї народо м на снѣденїе, то нариче тсꙗ жертва*). 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}, ирина {г}, еже е с мирнаꙗ*). 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 реко-мендаций TEI — Text Encoding Initiative (версия P5). На базе рекомендаций TEI уже были реализованы и продолжают функционировать более 150 проек-тов (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 и др., 2010; Захаров и др., 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 орга-низацией, что дает возможность пользоваться исключительно предложен-ными тегами и атрибутами, не усложняя код разметки дополнительными но-вовведениями.

Edited with the demo version of Infix Pro PDF Editor
To remove this notice, visit: www.iceni.com/unlock.htm

В качестве примера использования предлагаемых TEI тегов возьмем статью азбуковника Кида́ръ {ж} {ѱл҃о^м^ рѳі҃} (т) тма. и̓ли̓ клобукъ а̓рхїе̓ре́искїи. ї̓́же и̓ митра гл҃е/тсѧ. Ї па́ки кида^р^ нарица́е^т^ и̓ до́мы татар̔скїѧ. о̓вы бо ї̓́хъ в земли̓ и̓ско́паны, и̓ того рад‟ те́м̔ныи; о̓вы на колесни́ца^х^, по^л^стми̓ о̓гражде́ны; є̓́сть же и̓ гра^д^ нарица́емъ кида́ръ о̓ нем̔же пи́шеть, всели́хсѧ с селы кидаръскїми. Разметка данной статьи, ориентированная на структурную организацию, будет выглядеть следующим образом:

<entryFree>

<form><w>Кида́ръ<note><lang>ж</lang><bibl><title> ѱл҃ом рѳі҃</title></bibl> </note></w></form>

<sense> (т) <def n="1">тма.</def> и̓ли <def n="2"> клобукъ а̓рхїе̓ре́искїи. ї̓́же и̓ <mentioned>митра</mentioned> гл҃етсѧ.</def> Ї па́ки кидар нарица́ет и̓ <def n="3">до́мы татар̔скїѧ. о̓вы бо ї̓́хъ в земли̓ и̓ско́паны, и̓ того рад‟ те́мныи; о̓вы на колесни́цах, полстми̓ о̓гражде́ны;</def> <def n="4">є̓́сть же и̓ град нарица́емъ кида́ръ о̓ нем̔же пи́шеть, <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 фонети-ческие и графические приемы передачи иноязычных слов.

Список литературы

Бабалык и др., 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, 2005 — 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, 2006 — Вотинцев П.А. Использование формата TEI для об-мена данными с полнотекстовой информационно-поисковой системой «Манускрипт» // Современные 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, 2006. С. 30–31.

Захаров и др., 2011 — Захаров В.П., Митрофанова О.А., Михайлова В.Д. Разметка словарей в соответствии со стандартом TEI // Информацион-ные технологии в лексикографии. СПб., 2011. С. 61–69.

История, 2001 — История русской лексикографии / Отв. ред. Ф.П. Сороколетов. СПб., 2001.

Edited with the demo version of Infix Pro PDF Editor
To remove this notice, visit: www.iceni.com/unlock.htm