FILOLOGIJA 68, Zagreb 2017.

UDK 801.82:811.163:2
https://dx.doi.org/10.21857/yrvgqtkj39
Pregledni članak
Primljen 30.IV.2017.
Prihvaćen za tisak 19.II.2018.

Анна Марија Тотоманова
Софийски университет «Св. Климент Охридски»
Катедра по кирилометодиевистика
бул. «Цар Освободител» № 15, БГ-1540 София
atotomanova@abv.bg

ДИАХРОННЫЙ КОРПУС БОЛГАРСКОГО ЯЗЫКА: СОСТОЯНИЕ И ПЕРСПЕКТИВЫ

Доклад отражает историю создания Диахронного корпуса болгарского языка и электронных инструментов для обработки средневековых славянских текстов с целью создания веб базированного исторического словаря болгарского языка. В диахронный корпус входят тексты доказанно болгарского происхождения X—XVIII вв., принадлежащие к разным жанрам средневековой книжности. Корпус обладает своим собственным софтвером, который позволяет адекватное комментирование текста с палеографической, кодикологической и текстологической точки зрения. Тексты набраны специально разработанными для этой цели шрифтами по стандарту UTF-8. К настоящему моменту мы располагаем тремя шрифтами, которые инсталированы в конвертор, позволяющий превращение ранее набранных текстов в документы по новому стандарту. Сам корпус содержит свыше 130 текстов и постоянно пополняется новыми текстами по адресу *http://histdict.uni-sofia.bg/*. Перед каждым текстом опубликована информация об его источниках, датировке, издании, авторе и т.п. На том же сайте находится и полностью оцифрованный словарь древнеболгарского языка (*Старобългарски речник*), созданный Институтом болгарского языка при БАН. И корпус, и словарь находятся в свободном доступе, но потребителям видны только полностью отредактированные тексты.

Для разработки исторического словаря был создан специализированный софтвер для редактирования словарных статьей древнеболгарского словаря и создания новых словарных статей, так как исторический словарь разрабатывается на базе оцифрованного древнеболгарского словаря. К словарю разработана поисковая машина, которая облегчает работу по созданию новых словарных статьей. Ускоренным ходом идет работа и по созданию морфологического

аннотатора (таггера), прототип которого тоже расположен на сайте. Аннотатор разрабатывается с помощью созданного тагсета и грамматического словаря древнеболгарского языка, которые учитывают все возможные формы средневекового славянского языка разных изводов. И тагсет, и грамматический словарь опубликованы на сайте в свободном доступе.

Введение

Приношу свою благодарность организаторам за возможность участвовать в работе международной конференции, посвященной проблемам исторической лексикографии, и поделиться своим опытом в применении цифровых технологий в таких консервативных и традиционных областях, какими являются палеославистика и медиевистика. Мой доклад посвящен результатам и перспективам трех проектов, финансированных по программам Европейского социального фонда, в которых приняло участие более чем 80 аспирантов и молодых ученых до 35 лет и около 2 тысяч студентов. Первый проект BG051PO001-3.3-04-0011 Компьютерные и интерактивные инструменты для исторических языковедческих исследований начался в 2009 г. и ставил себе целью не только ускорить процесс сбора и обработки данных письменных памятников XI—XVIII вв., но и привлечь к исторической лингвистике молодых людей, для которых использование компьютера является частью их естественной среды и культуры[1].

Второй проект BG051PO001-3.3.06-0024 Информатика, грамматика, лексикография начался в 2012 г. и был задуман и осуществлен как продолжение первого проекта, который был признан лучшим проектом в схеме, предзназначеной для аспирантов/докторантов. Окончился второй проект в середине 2015 года[2].

Третий проект BG051PO001-4.3.04-0004 E-Medievalia. Электронные ресурсы для дистанционного обучения по медиевистике был выполнен в схеме, целью которой являлось создание электронных дистанционных курсов для студентов ВУЗ-ов. Он протекал одновременно со вторым проектом и окончился в конце 2014 года.

Подробную информацию об этих проектах можно найти на специализированном веб-портале Cyrillomethodiana *http://cyrillomethodiana.uni-sofia.bg*.

[1] О результатах этого проекта см. Тотоманова 2011 и Totomanova 2012.

[2] См. подробнее Тотоманова 2015.

Сн. 1. Портал Cyrillomethodiana

Начиная свою проектную деятельность, мы ставили перед собой две главные задачи. Первая состояла в создании софтверных инструментов для выработки исторического словаря болгарского языка, ввиду того что наш язык является первым священным и письменным языком славянства и соответственно обладает и самой продолжительной письменной историей. Вторая сосредотачивалась на оцифровке издания «Старобългарски речник», созданного сектором истории болгарского языка в ИБЕ при БАН[3]. Этот словарь должен был превратиться в базу для создания историческото словаря болгарского языка, посредством обогащения и пополнения его словарного состава с помощью материалов, входящих в диахронный электронный корпус средневековых и ранних новоболгарских текстов.

1. Стандарт исторического электронного словаря

Стандарт электронного словаря приобрел форму еще в ходе первого проекта, после того как в результате долгих обсуждений было принято решение о создании исторического словаря диахронного типа[4],

[3] Старобългарски речник. Т.I — II. С. 1999, 2009.

[4] Г. А. Богатова ввела и объяснила термины диахронный и синхронный исторический словари (Богатова 1981:83—84).

который должен представить историю болгарских слов со времени их первой письменной регистрации вплоть до сегодняшнего дня.

Исторический словарь такого типа по нашему мнению должен обладать, во-первых, широкими хронологическими границами (в нашем случае нижняя граница совпадает с началом письменного периода в IX в., а верхняя с нашей современностью). Во-вторых, он должен вырабатываться на базе неограниченного тематикой текстового корпуса, который включает в себя как книжные, так и некнижные тексты (топонимы, имена собственные, диалекты, разговорную речь, надписи и граффити). Словник электронного исторического словаря в принципе является открытым и обогащается и пополняется в ходе построения корпуса. Лексический материал в историческом словаре представляется в диахронном виде, что в свою очередь предполагает не только регистрацию разных значений слова, но и указания на их генетическую связь.

В текстовой корпус исторического словаря входят в первую очередь болгарские средневековые тексты: сочинения древнеболгарских писателей; переводы болгарского происхождения (сочинения Святых отцов, хроники, монашеская литература, историко-апокалиптические сочинения, юридические тексты, сборники устойчивого и смешанного состава и т.под. Во вторую он должен содержать и некнижные письменные памятники, которые включают записи писцов на полях книг, надписи и графитти, грамоты. Ранние новоболгарские памятники (прежде всего дамаскины и сборники смешанного характера), также как и диалектные тексты, тоже являются частью корпуса исторического словаря.

Методика разрабатывания электронной базы исторического словаря предполагает наличие двух основных баз-данных: оцифрованного «Старобългарски речник» и специализированного диахронного корпуса средневековых болгарских и ранних новоболгарских текстов. Для создания словаря будут использованы и другие, уже существующие специализированные корпуса, как *Български национален корпус*[5], *диалектные корпуса, Корпус на българската разговорна реч*[6] и под.

Поэтому еще в начале работы мы осознали, что создание баз-данных потребует соответствующего программного обеспечения и в целях ускоренной и надежной обработки текстов для выработки исторического словаря нужно будет разработать еще и целый набор электронных инструментов, начиная с поисковой машины, которая осу-

[5] См. описание и возможности пользования этого корпуса по *http://www.ibl.bas.bg/BGNC_bg.htm.*

[6] Корпус разработан в рамках инициативы BgSpeech и поддерживается Факультетом славянских филологий по адресу *http://bgspeech.net*.

ществляет поиск в словарях и в текстах корпуса; специализированного софтвера для редактирования словарных статей и создания новых словарных статей исторического словаря и специализированного софтвера для морфологической аннотации средневековых славянских текстов в широких хронологических границах с IX по XVIII в.

2. Специализированные древнеболгарские шрифты и конвертор

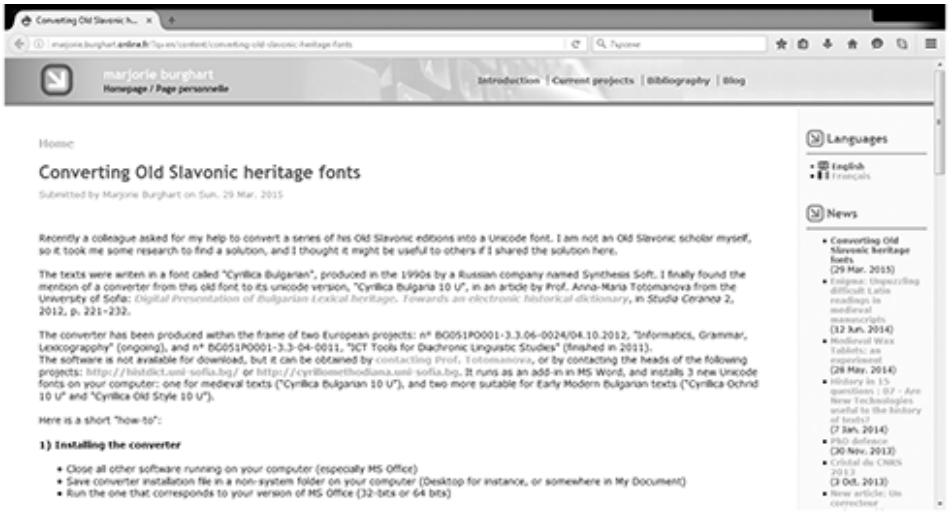
Перед началом работы мы нуждались прежде всего в специализированном древнеболгарском шрифте по стандарту Уникод с большим набором знаков для адекватной передачи особенностей славянского правописания разных изводов. К началу 2010 г. мы уже располагали одним таким шрифтом по стандарту UTF-8 *Cyrillica Bulgarian 10 U*, который был разработан по проекту «История и историзъм в православния славянски свят. Изследване на идеите за история», финансированному национальным научно-исследовательским фондом, и уже активно использовался для издания книг серии »История и книжнина« по этому проекту. В ходе новых проектов шрифт, который сопровождался конвертором для конвертирования ранее набранных на компьютере текстов, был усовершенствован и улучшен. Сегодня мы располагаем уже тремя древнеболгарскими шрифтами с разным дизайном: *CyrillicaBulgarian10U, CyrillicaOchrid10U* и *CyrillicaOldStyleU.*

Последний *(CyrillicaOldStyleU)* предназначен для набора ранних новоболгарских текстов, прежде всего дамаскинов и сборников смешанного содержания, и используется не только нами, но и коллегами в Канаде по проекту О. Младеновой *Pragmatic Function Words: A Corpus-Based Description of Variation*, который исследует историю балканских дамаскинов. Все шрифты хорошо работают под всякими редакторскими и издательскими программами. Чтобы прочитать документы в Интернете, не нужно инсталировать шрифт, и таким образом все набранное и выложенное нами в сети доступно всем со всех точек мира.

Сначала конвертор конвертировал документы, набранные всеми шрифтами семьи *CyrillicaBulgarian, CyrillicaOcrid* и *Cyrillica Shafarik*, разработанными компанией *Synthesis Soft*, которые широко использовались в Болгарии, но в ходе нашей работы сам конвертор был усовершенствован и к нему были добавлены новые возможности для конвертирования документов, набранных другими древнеболгарскими шрифтами как *PopRetkov*, используемым итальянскими славистами, и *Unicode* шрифт *BukyVede,* который тоже восходит к болгарским шрифтам, разработан-

ным комаиией *Synthesis Soft*. Конвертор превращает в уникодовский документ и греческие шрифты *TimesGreekClassic* и *TimesGreekOld*, разработанные той же компанией *Synthesis Soft*, и превращает все разновидности шрифта *TimesCyrillic* в *TimesNewRoman.* Последний, третий вариант конвертора разработан в двух версиях – 32- и 64-битовой в связи с появлением новых 64-битовых версий продуктов Microsoft.

Новые шрифты и конвертор получили распространение не только в Болгарии, но и за рубежом, и коллеги слависты, активно пользующиеся ими, создали специальный блог *Converting Old Slavonic heritage fonts*, содержащий информацию о фонтах, способах их приобретения и инсталяции на английском *http://marjorie.burghart.online.fr/?q=en/content/converting-old-slavonic-heritage-fonts* и на французском языках *http://marjorie.burghart.online.fr/?q=fr/content/convertir-des-textes-slaves-m-diaux-en-unicode.*



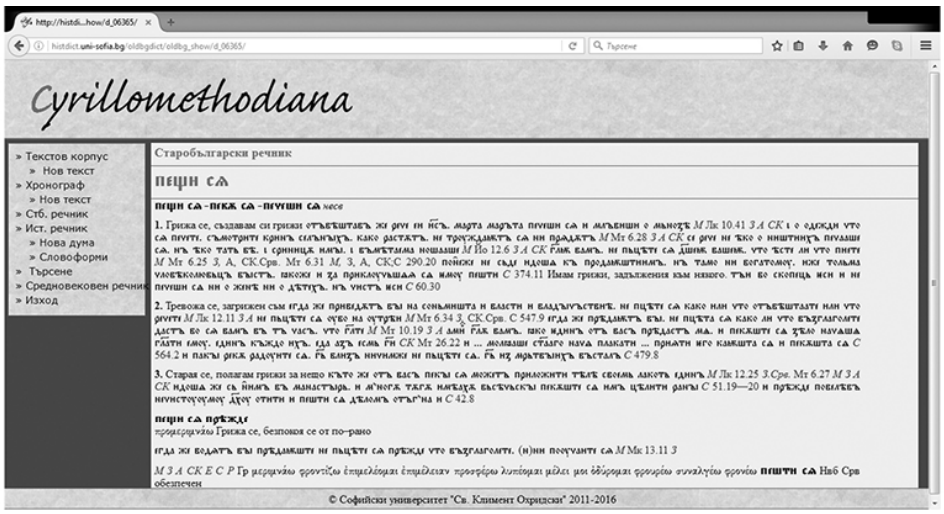
Сн. 2. Блог *Converting Old Slavonic heritage fonts*

За что мы им премного благодарны, так как у нас просто руки не доходили до этого, не смотря на то, что недавно я подарила конвертор издательству Oxford University Press по их просьбе и еще в 2014 г. разрешила издательству Brepols использовать шрифты для издания славянских синодиков в IV томе их серии COGD[7], который на днях выйдет в свет.

[7] *Conciliorum Oecumenicorum Generaliumque Decreta. I-VII. A Special Series of Corpus Christianorum Brepols,* 2006 — An International Research Program launched in Bologna and directed by † Giuseppe Alberigo and Alberto Melloni of FSCIRE, Fondazione per le Scienze Religiose Giovanni XXIII, Bologna.

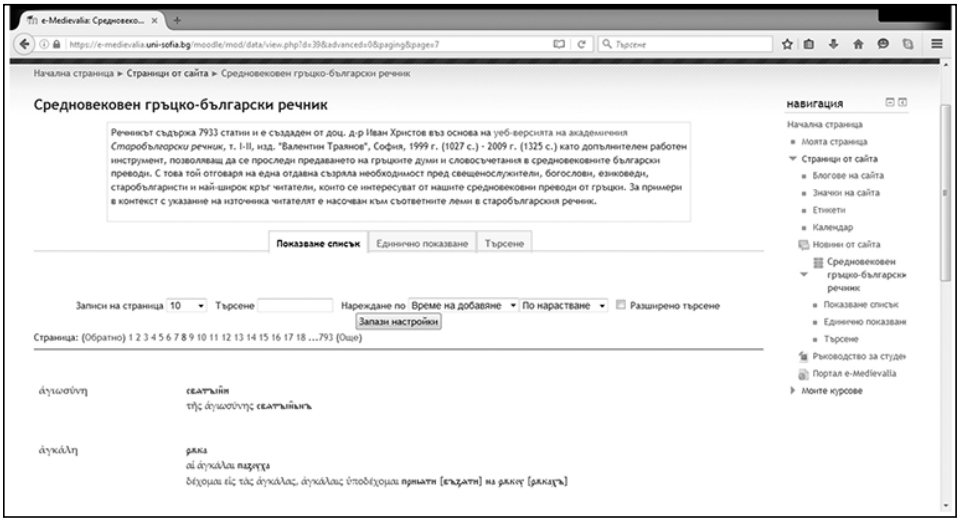
3. Диахронный корпус болгарского языка и оцифрованный словарь классического корпуса старославянского языка

Первый проект сосредоточился на создании корпуса средневековых и ранних новоболгарских текстов и на оцифровке двухтомного Старобългарски речник, в которых использовался шрифт *Cyrillica Bulgarian 10 U*. Для оцифровки словаря наш компьютерный специалист г-н Тодор Тодоров, который и является разработчиком шрифтов и конвертора, создал второй конвертор/генератор, который превратил словарь с его 11 000 статьями в структурированный XML документ, без потери важной информации. Первый вариант оцифрованного древнеболгарского словаря появился на сайте *histdict.uni-sofia.bg* в 2011 г. В 2014 у нас был уже усовершенствованный вариант словаря и только месяц назад к словарю были добавлены последние несколько сот лексем, которые из-за ошибок в оригинальном печатном издании не поддавались стандартизованной обработке. Так что теперь у нас полная цифровая копия древнеболгарского словаря, которая выложена в сети в свободном доступе.



Сн. 3. Оцифрованный словарь

На базе цифровой версии словаря наш коллега И. Христов, работающий по третьему проекту, сделал обратный греческо-древнеболгарский электронный словарь, который тоже можно найти на нашем сайте *histdict.uni-sofia.bg*.



Сн. 4. Обратный греческо-древнеболгарский электронный словарь

С гордостью можем сказать, что оцифрованный *Старобългарски речник* является первым полностью выложенным в сети лексикографическим пособием по палеославистике, а обратный электронный словарь вообще не имеет аналогов. Словарь оснащен также виртуальной клавиатурой, которая облегчает поиск в словарных статьях.

По тому же адресу *histdict.uni-sofia.bg* находится и диахронный текстовой корпус, который на данный момент содержит 130 текстов разной длины и разных жанров и постоянно пополняется новыми текстами. Для сравнения по завершении первого проекта в корпусе было только 75 текстов. В корпус входят средневековые славянские тексты болгарского происхождения и с разной орфографией (древнеболгарские, среднеболгарские, тырновские, ресавские, рашские, русские), новоболгарские дамаскины и записи средневековых книжников. В нем представлены не только переводные, но и оригинальные произведения древнеболгарских книжников в их жанровом многообразии – богослужебные, экзегетические, агиографические, юридические, хронографические, историко-апокалиптические тексты и др., причем некоторые из них не издавались до их публикации в корпусе. За проектный период, охватывающий 4 с половиной года, в корпус вошли сочинения древнеболгарских писателей — Климента Охридского, Йоанна Экзарха, Константина Преславского, Патриарха Евфимия, Константина Костенецкого; тексты Ефремовской кормчей, Хроники Константина Мана-

ссия, Диоптры Филиппа Монотропа; Пандекты Антиоха; перевод слов Йоанна Богослова, Германов сборник, Тиквешский и Берлинский сборники, Синодик царя Борила, несколько дамаскинов и другие важные для болгарской литературной и языковой истории памятники. Некоторые тексты были предоставлены нам коллегами из Италии и Канады. Тексты классического корпуса не вошли в корпус, так как этот период уже охвачен древнеболгарским словарем, который стал основой исторического словаря.

Софтвер корпуса можно описать как *user friendly* в настоящем смысле слова и он очень удобен для использования. Электронные инструменты для комментирования текстов как в палеографическом и кодикологическом аспектах (контекстуальные заметки), так и в текстологическом отношении (разночтения), создают новые возможности для адекватной и детальной передачи средневековых славянских текстов и уже использовались для цифрового издания всех текстов, входящих в Архивский хронограф.

Сн. 5. Диахронный корпус

Сам корпус является и прекрасным инструментом для цифрового представления болгарского лексического разнообразия в историческом плане. Открытость и доступность содержащихся в нем данных обеспечивает возможность его постоянного пополнения. Ввод текстов является максимально упрощенным, но может осуществляться только авторизованными пользователями. Для работающих над корпу-

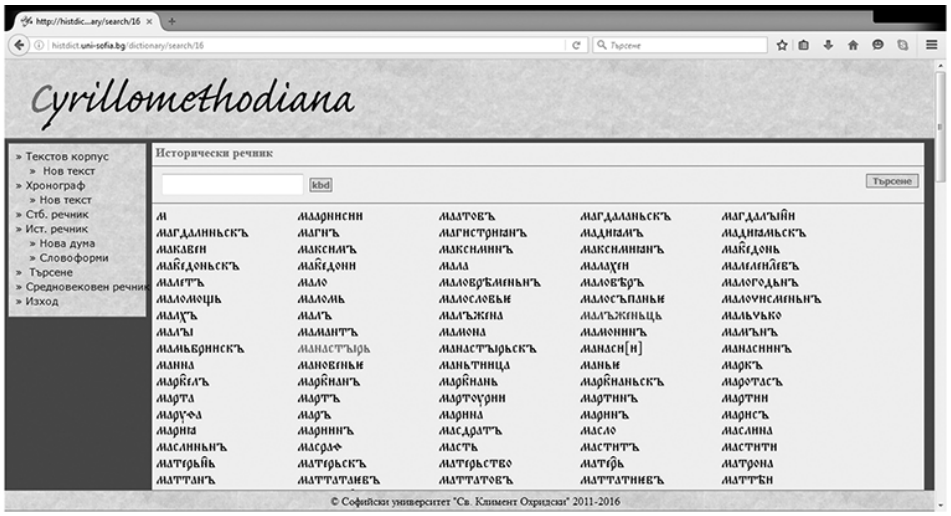
сом созданы разные уровни доступа, которые позволяют ввод новых текстов, редактирование старых и публикацию уже отредактированных текстов. Внешним пользователям видны только опубликованные тексты, прошедшие редакцию. Каждому тексту в корпусе предшествует стандартное описание с основными данными об его источнике, датировке, жанре и т. д.

4. Исторический словарь и поисковая машина

Второй проект не только продолжил работу по пополнению корпуса и по усовершенствованию его софтвера и улучшению работы конвертора, но сосредоточился на применении специализированного софтвера для исторического словаря, прототип которого был создан тоже по первому проекту. Задача оказалась довольно тяжелой, так как ситуация осложнялась обстоятельством, что использованная нами уже новая версия софтвера *Microsoft* оказалась несовместимой не только с софтвером словаря, но и с софтвером корпуса. Работа по устранению софтверных проблем отняла у нас больше года, но в конце концов структура корпуса была приведена в соответствие с новыми продуктами *Microsoft*, а работающий прототип словарного софтвера для создания и редактирования словарных статей появился уже в конце лета 2014 года. Несмотря на опоздание софтверного решения работа над созданием статей исторического словаря шла как было намечено в начале этого проектного предложения. На общем заседании участников проекта было принято решение, что новые словарные статьи будут создаваться не произвольным образом, а на тематическом принципе, и как объект лексикографической обработки была подобрана христианская терминология средневековых памятников. Выбор этой лексико-семантической группы не был случайным, так как мы надеялись с помощью созданных инструментов дополнить наблюдения Ф. Миклошича, сделанные почти полтора века тому назад (Miklosich 1876). К тому же создание терминологического аппарата, связанного с принятием христианства и отправлением культа, является основополагающим для развития и функционирования древнеболгарского языка как священного и литературного языка православных славян[8]. На основе исследования Миклошича с привлечением данных новых словарей-индексов мы создали словник лексикографического исследования, в который вошло около 500 лексем. Кроме того было принято решение, что для создания статей исторического словаря мы будем опираться не только на

[8] Об этом см. Илиева 2013.

данные корпуса, но и на данные всех существующих лексикографических пособий, опубликованных до 2012 г. С этой целью они были отсняты цифровым способом и копии были предоставлены авторам словарных статей. До появления работающего словарного софтвера статьи создавались в doc формате и только после этого и надлежащего обучения авторов новые словарные статьи вошли в структуру электронного исторического словаря, который тоже расположен по адресу *http://histdict.uni-sofia.bg*.

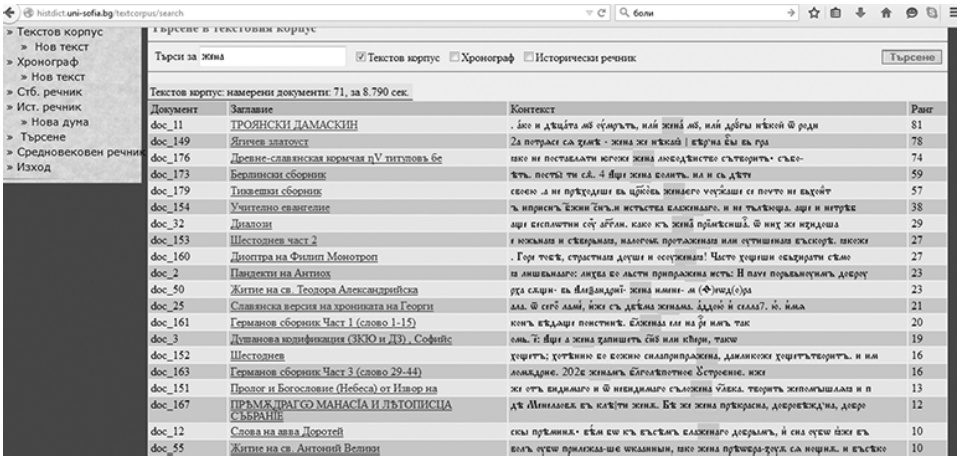


Сн. 6. Исторический словарь

К концу проекта мы уже перевыполнили наши планы по созданию и редактированию словарных статей и их количество превысило первоначально намеченные 300 статей. В списке лексем под каждой буквой новые и редактированные словарные статьи появляются в цвете.

Одновременно с созданием статей нового словаря шла работа и по созданию поисковой машины, которая должна была облегчить работу по словарю. К сожалению, первый прототип этой машины, созданный по системе CLaRK, оказался тупиковым решением и пришлось начинать сначала и приступить к созданию поисковой машины, работающей в системе histdict. Новая машина поиска является надежной, но нуждается в усовершенствовании, так как короткие сроки для ее создания не позволили добавление дополнительных функций. Она показывает список текстов, содержащих искомую форму, но сами словоформы показываются только в конкретном тексте корпуса при использова-

нии возможностей внутреннего поиска браузера. Функция поиска по корпусу тоже пока недоступна внешним пользователям.



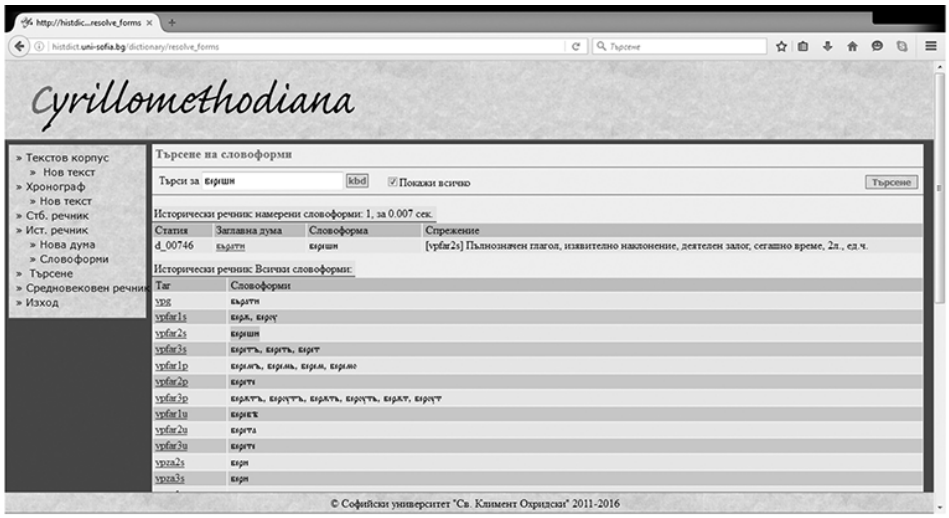
Сн. 7. Поисковая машина в системе *histdict.uni-sofia.bg*

Искомый стринг вводится с помощью виртуальной клавиатуры словаря:

Сн. 8. Виртуальная клавиатура

5. Морфологическая аннотация средневековых болгарских текстов

По ходу работы мы пришли к пониманию, что создание быстро работающей и надежной поисковой машины тесно связано с разрабатыванием надежной системы для морфологической аннотации средневековых текстов. По настоянию софтверных специалистов вместе с моими коллегами Т. Славовой и Г. Ганевой я включилась в создание набора тагов (тагсет) для всех возможных словоформ в средневековых болгарских текстов. Тагсет создавался на базе существовавшего тагсета для современного болгарского языка, который был дефинирован и использовался в синтаксическом корпусе *BulTreeBank*, в котором 215 149 слов (Simov, Osenova, Slavcheva 2004). Составленный нами тагсет состоит из 2 200 тагов/этикеток с учетом разных фонетических вариантов конца слова в средневековых славянских памятниках разных редакций (Тотоманова, Славова, Ганева 2015). Вслед за тагсетом мы разработали и грамматический словарь средневекового болгарского языка, который состоит из таблиц, содержащих все словоизменительные парадигмы имен, местоимений и глаголов. Сам тагсет и принципы его изготовления, а также и грамматический словарь, опубликованы и на сайте *http://cyrillomethodiana.uni-sofia.bg.*



Сн. 9. Морфологический аннотатор

Используя тагсет и грамматический словарь, Тодор Тодоров успел к середине прошлого года создать и запустить прототип морфологического аннотатора, который находится в пространстве исторического словаря *http://histdict.uni-sofia.bg/dictionary/resolve_forms*, и постоянно пополняется новыми формами.

6. Применение электронного корпуса и электронных инструментов

Диахронный корпус болгарского языка и разработанные электронные инструменты можно использовать не только для создания электронно базированных лексикографических пособий разного типа как диахронных исторических словарей, исторических словарей синхронного типа (словари книжности определенного исторического периода, словари языка отдельного книжника или книжного центра), словарей-индексов отдельных памятников[9], тематических словарей, этимологических (историко-этимологических) словарей и обратных словарей, но и в научных целях для проведения исторических лингвистических исследований во всех областях лингвистики (в морфологии и морфосинтаксисе, морфонологии, в фонетике и фонологии, лексикологии, этимологии, дериватологии, фразеологии, текстологии и орфографии). На базе корпуса уже создано несколько докторских диссертаций и написано множество научных сообщений и статей.

Тексты корпуса могут послужить основой и для изготовления электронных и бумажных изданий средневековых письменных памятников, и для создания лексикографических и учебных пособий, таких как хрестоматий, словарей, учебников.

Корпус оказался и прекрасным инструментом для популяризации болгарского культурного и книжного наследия в стране и за рубежом. В будущем мы намерены опубликовать параллельные переводы средневековых текстов на современном болгарском и при возможности и на других европейских языках, чтобы расширить круг пользователей корпуса и его применения в целях распространения информации о богатом культурном и книжном наследии нашего народа.

[9] Уже разработан и издан словарь-индекс Синодика царя Борила (Тотоманова, Христов 2015).

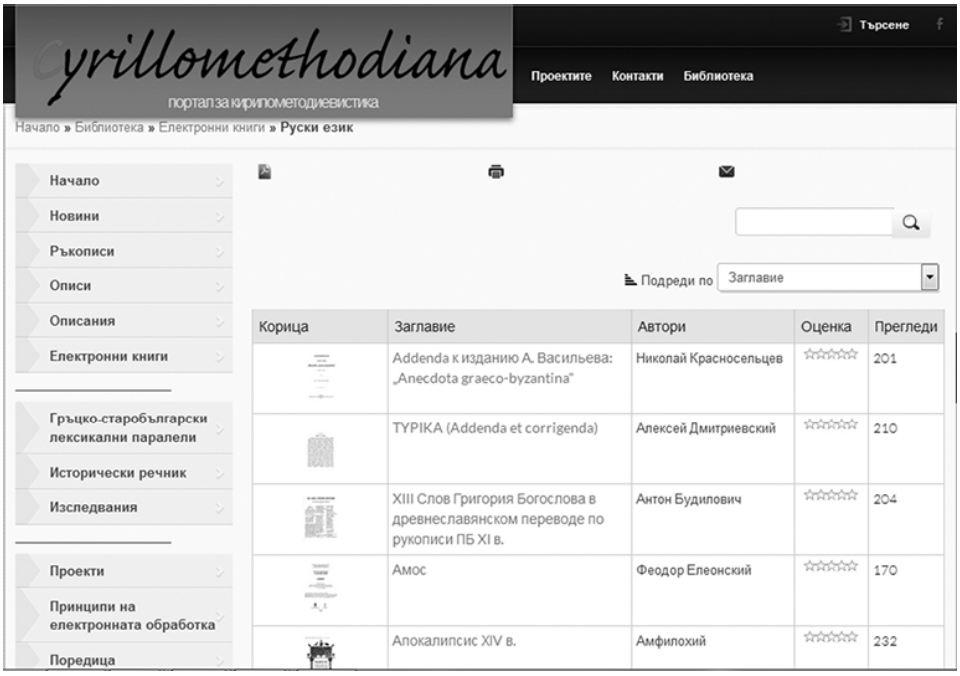
7. Проект *e-Medievalia* и его связь с системой *Histdict*

Еще в начале 2012 года мы решили использовать накопленный опыт по созданию электронных инструментов для исследований в области палеославистики и в целях преподавания медиевистики и написали проект о создании специального веб-портала, который получил название e-Medievalia. В очень короткий срок (24 месяца) коллектив проекта, который состоял из 50 специалистов разного профиля – филологов, историков, философов, богословов, историков искусства, инженеров-информатиков – успел создать уникальную веб-базированную систему для дистанционного обучения по медиевистике, которая дополняет обучение по всем медиевистическим дисциплинам. Портал содержит 24 курса, представляющие собой цельную и уникальную программу, вбирающую в себя почти все медиевистические дисциплины и организованную в четыре основных курса и восемь модулов. Содержание курсов отличается высокой степенью интерактивности – они включают лекции, тестовые задачи, связи между отдельными учебными единицами и разными курсами, изображения, практикумы для самостоятельной подготовки, ссылки на другие электронные ресурсы,

Сн. 10. *E-Medievalia*

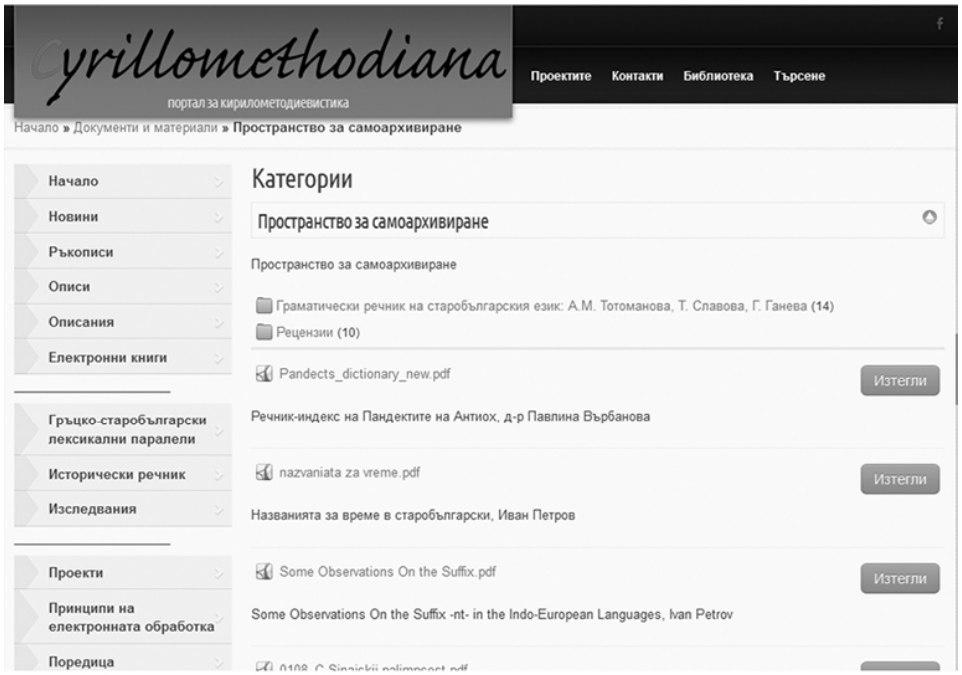
обратную связь, а не просто картинки, которые можно скачать. Основные курсы — древнеболгарский (старославянский) язык, история болгарского языка и древнеболгарская литература — доступны и на английском языке. Свыше 2000 студентов прошло через систему за два с половиной года и она продолжает активно использоваться. Наши будущие планы включают ее дополнение и перевод остальных курсов и ресурсов на английский язык, а также и привлечение коллег и студентов из других европейских университетов как авторов новых курсов и пользователей учебного веб-портала. Список курсов и их авторов вместе с краткими аннотациями самих курсов можно увидеть по адресу e-medievalia.uni-sofia.bg. Само учебное содержание однако доступно только после получения пароля от администраторов сайта.

Студенты могут пользоваться и электронной библиотекой с работами по палеославистике и медиевистике, которая предлагает больше 600 электронных книг на разных языках. Ее можно найти по адресу *http://cyrillomethodiana.uni-sofia.bg/index.php/component/booklibrary/218/all_category?Itemid=218*, и она является одной из самых посещаемых рубрик нашего портала C*yrillomethodiana*.



Сн. 11. *Электронная библиотека*

На том же сайте с прошлого года функционирует рубрика для электронных публикаций, в которой участники проектов, докторанты и студенты могут публиковать свои труды. Она называется Пространство для самоархивирования и видна по адресу *http://cyrillomethodiana.uni-sofia.bg/mdocs/category/8-archive*. Документы электронной библиотеки и электронные публикации можно скачивать по указанным ссылкам.



Сн. 12. *Архивирование*

8. Поддержка и развитие созданных электронных ресурсов

Так как проектное финансирование кончилось еще в 2015 году, Софийский университет выделил средства на поддержку системы *histdict* и веб-платформы *e-Medevalia* до конца 2017 года. В 2016 году мы заново отредактировали грамматический словарь и приписали леммам исторического словаря определенный тип словоизменения согласно их грамматическим характеристикам. Теперь нажимая на таг +, которым помечены изменяемые слова, можно увидеть всю парадигму данной лексемы в разных орфографических и морфологических вариантах:

Сн. 13. Электронный грамматический словарь

До сих пор наши усилия были направлены главным образом на создание и испытания электронных ресурсов и инструментов, так что настоящая и долгая работа по созданию и пополнению исторического словаря болгарского языка еще предстоит. В 2017 г. коллектив сосредоточится на создании програмы для редактирования электронного грамматического словаря, которая позволит исправлять и устранять ошибки online и создавать новые правила по ходу поступления словарных единиц. Это подводит нас еще ближе к автоматизации морфологического аннотатора и усовершенствованию поисковой машины. Пока что разрабатывание электронных инструментов и пополнение баз данных отнимало у нас большую часть времени и почти не оставалось времени на распространение наших результатов, чтобырасширить применение электронных инструментов в палеославистических исследованиях и облегчить доступ к электронным ресурсам и курсам.

В заключение могу сказать, что хотя мы начали работу только семь лет назад, все-таки мы сумели доказать, что палеославистика и новые технологии совместимы.

Литература

Богатова, Галина Александрова. 1981. *Историческая лексикография как жанр*. *Вопросы языкознания*, 1. 80—89.

Илиева, Татяна. 2013. *Терминологичната лексика в Йоан-Екзарховия превод «De fide orthodoxa»*. Печатница «Славейков»: София.

Miklosich, Franz. 1876. *Die christliche Terminologie der Slavischen Sprachen:* Eine sprachgeschichtliche Untersuchung von Franz Miklosich. Denkschriften der kaiserlichen Akademie der Wissenschaften. Philosophisch-Historische Klasse. Band 24. Wien, 1876.

Simov, Kiril, Petya Osenova, Milena Slavcheva. 2004. BTB-TR03: BulTreeBank Morphosyntactic Tagset, 2004. URL: *http://www.bultreebank.org/TechRep/BTB-TR03.pdf* (29.01.2017).

Totomanova, Anna-Maria. 2012. Digital Presentation of Bulgarian Lexical Heritage. Towards an Electronic Dictionary. *Studia Ceranea* 2, 221—234.

Тотоманова, Анна-Мария. 2011. Проектът «Компютърни и интерактивни средства за исторически езиковедски изследвания» и дигиталното представяне на словното богатство на българския език през вековете. Проект «Компютърни и интерактивни средства за исторически езиковедски изследвания». *Сборник доклади от заключителната конференция* 15.12.2011. ПАМ Пъблишинг Къмпани: София. 5—15.

Тотоманова, Анна-Мария, Татяна Славова, Гергана Ганева. 2015. Морфосинтактичен тагсет на старобългарския книжовен език. Информатика, граматика, лексикография *BG051-3.3-06-0024/2012. Сборник доклади и материали от заключителната конференция, София, 29—30.06. 2015 г.* ПАМ Пъблишинг Къмпани: София. 5—16.

Тотоманова, Анна-Мария, Иван Христов. 2015. *Речник-индекс на словоформите в Бориловия синодик и придружаващите го текстове в ръкопис НБКМ 289*. ПАМ Пъблишинг Къмпани: София.

Dijakronijski korpus bugarskoga jezika: trenutno stanje i perspektive

Sažetak

U članku se iznosi povijest stvaranja *Dijakronijskoga korpusa bugarskoga jezika* i digitalnih alata za obradu srednjovjekovnih crkvenoslavenskih tekstova potrebnih za izradu *Povijesnoga rječnika bugarskoga jezika na svemrežju*. Dijakronijski korpus uključuje tekstove različitih žanrova kojima je dokazano bugarsko podrijetlo. Korpus je zasnovan na vlastitom programu koji omogućuje primjereno komentiranje s paleografske, kodikološke i tekstološke točke gledišta. Tekstovi su digitalno tipizirani s pomoću posebno konstruiranih starocrkvenoslavenskih UTF fontova. Trenutno imamo na raspolaganju tri fonta i pretvarač koji prethodno tipizirane tekstove koji nisu u *Unicodeu* prenosi u dokumente u *Unicodeu*. Do sada je u korpusu objavljeno više od 130 tekstova, a još ih je u postupku pripreme za prijenos na mrežne stranice. Korpus se nalazi na: *http://histdict.uni-sofia.bg*. Svaki je tekst uveden rubrikom koja sadrži podatke o njegovu izvoru, dataciji, izdanju, autoru ili autorima itd. Mrežne stranice uključuju i potpuno digitaliziranu inačicu *Starocrkvenoslavenskoga rječnika* (bug. Старобългарски речник), učinjenoga u *Institutu za bugarski jezik* BAS-a. Oboje, korpus i rječnik, objavljeni su kao izvori u otvorenom pristupu, s tim da je korisnicima dopušteno vidjeti samo potpuno uređene tekstove.

Poseban je program stvoren za potrebe autora *Povijesnoga rječnika bugarskoga jezika*. S obzirom na činjenicu da je taj povijesni rječnik zasnovan na digitalnoj inačici *Starocrkvenoslavenskoga rječnika*, program omogućuje uređivanje postojećih natuknica i stvaranje novih. Također, izrađena je tražilica kojoj je svrha olakšati rad na novom rječniku. Nedavno smo se usredotočili na stvaranje morfološkoga označivača (eng. *tagger*), čiji je prototip također dostupan na mrežnim stranicama. Morfološki je označivač zasnovan na punom rasponu morfoloških oznaka te na gramatičkom rječniku srednjobugarskoga rječnika. Puni raspon morfoloških oznaka i gramatički rječnik također su dostupni na mrežnim stranicama, a zajednički daju potpuni opis svih oblika u srednjocrkvenoslavenskim tekstovima.

Ključne riječi: dijakronijski korpus bugarskoga jezika, tražilica, specijalizirani program za *Povijesni rječnik bugarskoga jezika*, morfološko označivanje, gramatički rječnik srednjovjekovnoga bugarskoga

Ключевые слова: *Диахронный корпус болгарского языка*, поисковая машина, специализированный софтвер *для* словарь древнеболгарского языка (*Старобългарски речник*), морфологический аннотатор