А. Тотоманова
СУ „Свети Климент Охридски“

ПРОЕКТЪТ „ИНФОРМАТИКА, ГРАМАТИКА, ЛЕКСИКОГРАФИЯ“ И ДИГИТАЛНАТА ОБРАБОТКА НА СРЕДНОВЕКОВНИ СЛАВЯНСКИ ТЕКСТОВЕ

Проектът BG051PO001-3.3.06-0024 „Информатика, граматика, лексикография“ беше замислен и структуриран като продължение на предишния успешен докторантски проект, финансиран от ЕСФ по ОП РЧП BG051PO001-3.3-04-0011 „Компютърни и интерактивни средства за исторически езиковедски изследвания“, който беше определен за най-добрия проект в схемата „**Подкрепа за развитието на докторанти, постдокторанти, специализанти и млади учени“**[1]. Затова още от самото си начало новият проект започна да използва електронните ресурси и инструменти, изработени по предходния проект, като се стараеше да ги усъвършенства, да обогати възможностите им и да ги допълни с нови. Както и в предходния проект целевата група от докторанти, специализанти и млади учени беше въвлечена в изпробването и използването на разработваните електронни инструменти както за нуждите на самия проект, така и в собствената си изследователска практика. И докато първият проект се съсредоточи върху изработването на диахронния исторически корпус, то фокусът на този проект беше поставен върху разработването на ***електронния исторически речник на българския език*** и върху създаването на ***специализиран софтуер за морфологична анотация на средновековни славянски текстове в широки хронологически граници от IX до XVIII в.,*** като попълването на корпуса с нови дигитално набрани текстове се превърна вече в рутинна дейност.

По първата задача още в първите месеци на проекта беше взето решение новите речникови статии от историческия речник да бъдат подбирани по тематичен признак и като обект на лексикографската обработка беше избрана християнската терминология в средновековните паметници. Основанията за избора на тази група лексеми бяха следните:

1. Изграждането на терминологичен апарат, свързан с изповядването на християнството и обслужването на култа е въпрос от първостепенна важност за развитието и функционирането на старобългарския език като свещен и книжовен език на православните славяни.
2. Единственото обзорно и систематично изследване на християнската терминология в славянските писмени паметници „Die christliche Terminologie der Slavischen Sprachen“ принадлежи на Фр. Миклошич

[1] За резултатите от този проект вж. *Тотоманова 2011* и *Totomanova 2012*.

и отразява състоянието на познанията за лексикалното богатство на старобългарския книжовен език по времето на своето създаване, т.е. към 1876 г. , която е годината на издаването на текста, както и възможностите на тогавашната езиковедска методология.

3. Тематичният подход към лексиката на българския език в диахронен план би позволил след обработката на отделните лексико-семантични групи да се направят изводи за пътищата на интелектуализация на езика през Средновековието и за развитието на книжовната лексика.

И тъй като вече разполагахме с корпус от средновековни славянски текстове, който постоянно се попълва и разширява, с дигитализиран старобългарски речник, който напълно отразява състоянието на класическия корпус текстове, многобройни лексикографски наръчници, които могат да се ползват в дигитализиран или поне частично дигитализиран вид, ние решихме, че с помощта на новите технологии можем да допълним и разширим наблюденията на Миклошич, като държим сметка и за променената методология на изследването на този лексико-семантичен клас[2].

Започнахме с попълването на словника на бъдещото лексикографско проучване, като за целта списъкът с християнски термини по Миклошич беше допълнен с нови лексеми, ексцерпирани от достъпните ни християнски наръчници. Получи се един провизорен списък от около 500 лексеми, които бяха разделени между участниците в екипа, натоварени с изработването на статии за историческия речник.

Още тогава възникна въпросът дали при изработването на новите статии да се използват само данните от корпуса, който е отворен и непрекъснато се попълва, или да се включват и лексеми, вече включени в други лексикографски наръчници (речници и речник-индекси). Везните надделяха в полза на второто решение и започна събирането на изданията на излезлите до началото на 2012 г. лексикографски помагала, които бяха дигитализирани и разпратени на авторите на речникови статии . Работата по издирването и дигитализацията им отне няколко месеца и ни се стори, че същинската работа по историческия речник вече може да започне.

Но тук започнаха да изникват редица софтуерни проблеми. На първо място се оказа, че разработеният по предишния проект специализиран софтуер за редактиране на речникови статии от дигитализирания Старобългарски речник и изработване на нови не работи достатъчно добре и трябваше да се отстранят проблемите, които пречеха на пълноценното му използване. Това се оказа много тежка и трудоемка задача, като ситуацията беше усложнена допълнително от факта, че използваната от нас нова версия на софтуера на Microsoft за около две години беше станала несъвместима със софтуера както на речника, така и на текстовия корпус,

[2] По този въпрос вж. *Илиева 2013*.

който трябваше редовно да се попълва с новонабрани или конвертирани средновековни текстове. Тези обстоятелства станаха известни на екипа на проекта по време на двата семинара за изработване на електронни речникови статии и за изработване на исторически електронни корпуси на 20.04.2013 г. Отстраняването на софтуерните проблеми отне повече от година, но в крайна сметка софтуерната структура на корпуса беше обновена и приведена в съответствие с новите продукти на Microsoft, а работещ прототип на софтуера за изработване и редактиране на речникови статии се появи едва в края на лятото на 2014 г. Софтуерът беше представен на колегиума на целевата група, проведен на 27.09.2014 г. и беше придружен с нова шеста версия на специализираните старобългарски шрифтове и конвертора, към който бяха добавени нови шрифтове и функционалности.

По този начин в момента разполагаме с три старобългарски шрифта:

1. ***CyrillicaBulgarian10U***
2. ***CyrillicaOchrid10U***
3. ***CyrillicaOldStyle10U***

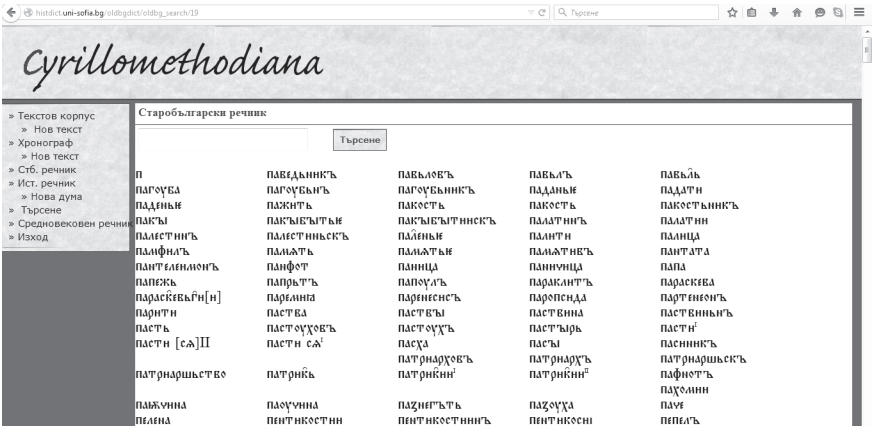
Последният ***(CyrillicaOldStyle10U)*** е предназначен за набор на ранни новобългарски текстове, предимно дамаскини и сборници с различно съдържание, и освен за нуждите на нашите проекти се използва и в проекта на О. Младенова *Pragmatic Function Words: A Corpus-Based Description of Variation* за изследване на балканските дамаскини, разработван в университета в Калгари, Канада. В хода на работата ни самият конвертор беше подобрен и към него бяха добавени нови нови функционалности:

- конвертиране на документи, набрани със старобългарския шрифт ***PopRetkov***, използван от италианските слависти, към новите Unicode шрифтове,
- конвертиране на гръцките фонтове ***TimesGreekClassic*** и ***TimesGreekOld***, също разработени от *Synthesis Soft*, към ***Palatino,***
- конвертиране на всички разновидности на шрифта ***TimesCyrillic*** *към* ***TimesNewRoman***.

При това новият вариант на конвертора е разработен в две версии – 32- и 64-битова с оглед на появата на нови 64-битови версии на продуктите на Microsoft.

Още един страничен продукт от усъвършенстването на шрифтовете, конвертора и речникарския софтуер беше появата на нова версия на дигитализирания през 2011 г. Старобългарски речник. Новата версия на дигитализирания речник също беше представена колегиума в края на септември и качена на сървъра на проекта. В нея са отстранени всички грешки на първата версия и са добавени 500 речникови статии, които поради технически пречки не бяха включени в предходната версия. Допълненият и поправен дигитализиран вариант на Старобългарския речник може да се види и използва напълно свободно на адрес http://histdict.

uni-sofia.bg в рубриката Старобългарски речник. На същия адрес беше създадена още една рубрика – исторически речник, където словникът на старобългарския речник може да се допълва с нови лексеми и речникови статии, а съществуващите речникови статии могат да бъдат редактирани и обогатявани с нови употреби и примери.



Фиг. 1. Нова версия на дигитализирания Старобългарски речник

В хода на сътрудничеството ни с проекта e-Medievalia, по който изработихме обучителна платформа с курсове по медиевистика, на базата на дигитализирания старобългарски речник колегата Н. Христов, който участва и в двата проекта, изработи гръко-старобългарски речник, към който ни отвежда линк, озаглавен Средновековен речник, който е незаменимо помагало за всички, които се занимават с техниката на превода през Средновековието.

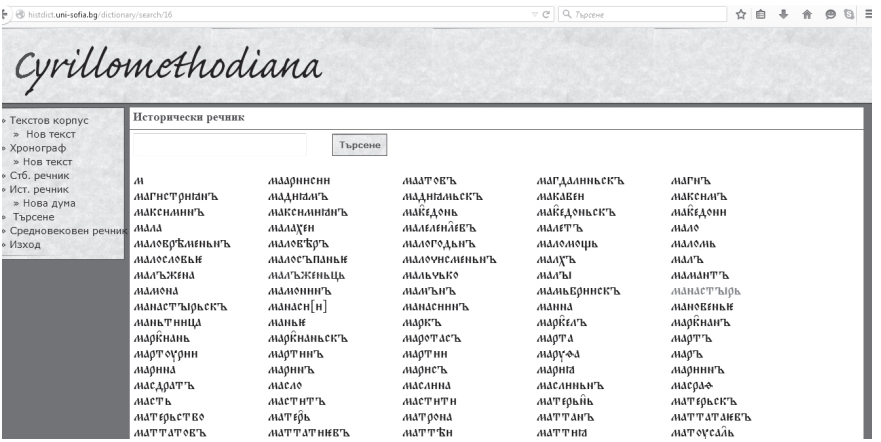
Друга група проблеми, свързани с изработката на речникови статии за историческия речник, възникна поради обстоятелството , че диахронният корпус със средновековни и ранни новобългарски текстове не разполагаше със собствена търсачка, която да позволява търсене във всички текстове едновременно, а трябваше да се използва търсачката на браузъра, за да се търси във всеки един текст поотделно, което усложняваше работата и удължаваше времето за търсене.

При така създалата се в началото на проекта ситуация – без надежден софтуер и търсачка, групата от експерти – научни ръководители, специализанти и постдокторанти, чиято задача беше създаване на речникови статии за историческия речник, реши да започне работа, като създава речниковите статии в Word формат, използвайки разработените вече старобългарски шрифтове по стандарта Unicode. За улеснение експертите доц. К. Симов и доц. П. Осенова, които трябваше да се заемат с изработ-

ването на морфологичен анотатор (тагер), предложиха да използват системата на CLaRK за търсене в езикови корпуси, за да изготвят конкорданс с всички появи на интересуващите ни лексеми. В тази връзка експертите речникари модифицираха провизорния списък от думи за нуждите на търсенето (нов списък), след което първият вариант на конкорданса беше разпратен на авторите на речниковите статии. Освен интересуващите ни словоформи на лексемите от словника той съдържаше и достатъчно голям контекст за определянето на значението на думата, както и номера на документа, от който е ексцерпирана информацията. Тъй като корпусът се попълваше непрекъснато с нови текстове по време на проекта, през м. март 2014 г. беше изготвен нов вариант на конкорданса (конкорданс II) и също разпратен на авторите на речниковите статии.

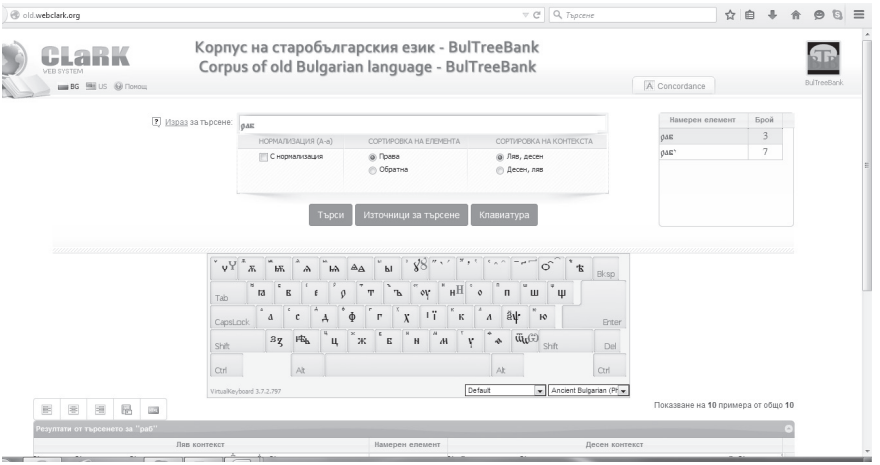
За нуждите на историческия речник беше изработен и утвърден списък със съкращения на изворите, използвани при работата над речниковите статии, който също беше разпратен на речникарите.

Както вече споменах, новият работещ софтуер беше представен на колегиума на целевата група, проведен на 27.09.2014 г. , и обсъден с експертите и авторите на статии, като едва тогава неговото активно използване стана възможно. Той работи върху новата версия на дигитализирания старобългарски речник, като за нуждите на историческия речник беше създадено отделно web пространство, озаглавено ***Исторически речник***, в което се изпробва и ползва лексикографският софтуер. Променените речникови статии се показват с различен цвят – в синьо новите, а в зелено редактираните статии. Експертите лексикографи през цялото време работят в близко сътрудничество със софтуерния специалист Тодор Тодоров, като му докладват за проблемите при използването на софтуера, които той постепенно отстранява.



Фиг. 2. Исторически речник

По същото време се появи и прототип на търсачката, базиран върху системата CLaRK, който би следвало да позволява търсенето както в отделни текстове по избор, така и в целия диахронен корпус, вж. http://old.webclark.org, снабдена с виртуална клавиатура, създадена от Тодор Тодоров.

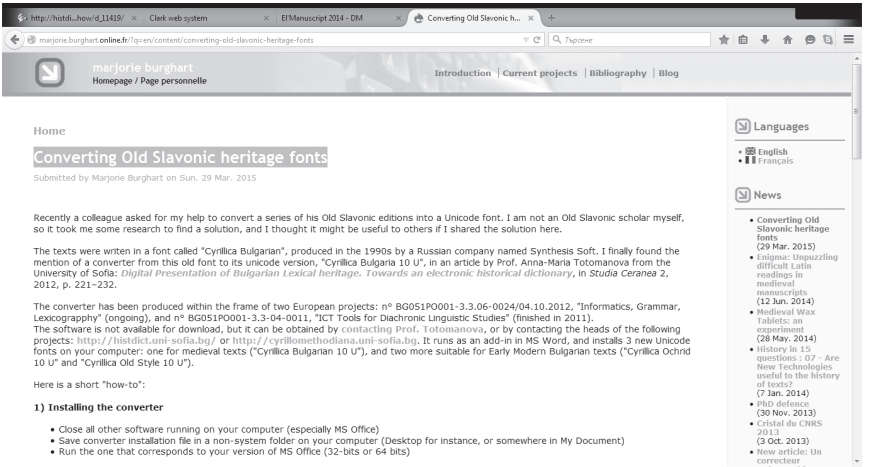


Фиг. 3. Търсачка в системата CLaRK

Всички електронни инструменти, разработени и усъвършенствани по проекта, бяха представени на Петата международна конференция El' Manuscript „Textual Heritage and Information Technologies" (http://textualheritage.org), която се проведе в курортния комплекс Камчия от 15 до 20.09.2014 г.

Тук без излишна скромност ще отбележа, че резултатите от двата ни докторантски проекта превъзхождат резултатите на подобни инициативи, които в други страни се изпълняват от цели институти. И заслугата за това е на целия екип на проекта и на неговата целева група, за което благодаря на всички активни участници в проектните дейности и особено на г-н Тодор Тодоров, който отговаряше и отговаря на всички наши софтуерни капризи.

Струва си да се добави също така, че шрифтовете и конверторът получиха разпространение не само у нас, но и в чужбина, и колеги слависти, които активно ги използват, направиха специален блог ***Converting Old Slavonic heritage fonts*** с информация за фонтовете и начина на получаването и инсталирането им на английски http://marjorie.burghart.online.fr/?q=en/content/converting-old-slavonic-heritage-fonts и на френски език http://marjorie.burghart.online.fr/?q=fr/content/convertir-des-textes-slaves-m-di-aux-en-unicode.

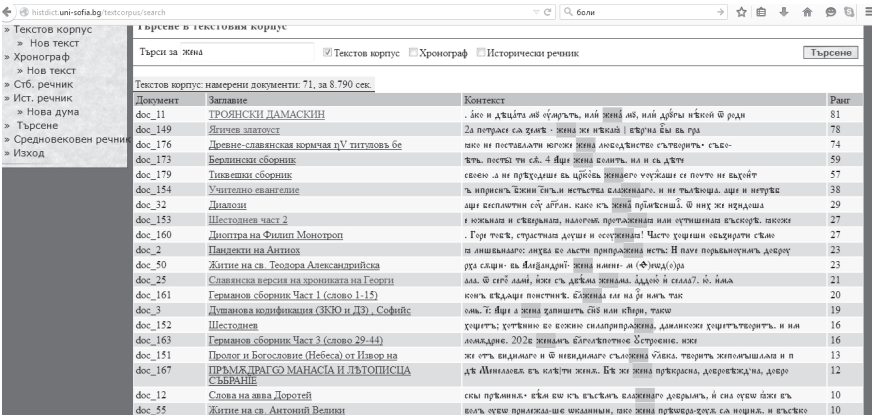


Фиг. 4. Блог за шрифтовете и конвертора

Нещо, за което на нас просто не ни достига време.

Изпробването на търсещата машина по системата CLaRK обаче даде разочароващи резултати още от самото начало. Търсенето беше бавно, като тежкият и морално остарял софтуер затормозяваше и работата на останалите електронни инструменти, разположени на сървъра на проекта. В редица случаи операцията завършваше с отговор, че е възникнала грешка. Когато все пак се получаваха резултати, те очевидно бяха непълни, което беше проверено многократно с общоупотребителни думи, за които търсачката ни посочваше само една или две появи. Окончателно се убедихме, че изработването на търсачка по системата CLaRK е задънена улица, след като благодарение на една кратка визита в Испания по програма COST, осъществена от д-р Гергана Ганева, се запознахме с опита на колегите, които от няколко десетилетия работят върху историческия речник на испанския език. Те споделиха, че преди години също са се опитвали да изработят търсачка, базирана върху този софтуер, и са имали същите проблеми, което в крайна сметка ги е накарало да изоставят този софтуерен подход към търсещата машина. И тук отново на помощ ни се притече г-н Тодор Тодоров, който разшири функционалностите на системата http://histdict.uni-sofia.bg с опцията *търсене,* която позволява да се търси не само в корпуса, но и в *Архивския хронограф* и в *Историческия речник*. Тази нова търсачка е много по-бърза и надеждна от предишната, но има нужда от доработване и усъвършенстване, затова засега е достъпна само за вътрешни потребители. В момента тя посочва паметниците, в които се среща написаният в прозорчето стринг и честотата на употребите му във всеки един от тях. След което с кликване върху заглавието на

паметника може да се търси чрез браузъра вътре в самия текст, като по този начин се открояват всички употреби.



Фиг. 5. Търсачка в системата http://histdict.uni-sofia.bg

Покрай одисеята с търсачката започнахме да осъзнаваме сложността на задачата, с която сме се нагърбили, и огромната отговорност, която сме стоварили върху плещите на софтуерните специалисти, тъй като такива електронни инструменти върху средновековен, при това ненормализиран, славянски материал се изработваха за пръв път. Постепенно както ние, така и софтуерните експерти, стигнахме до разбирането, че разработването на бърза и надеждна търсачка е свързано с разработването на надеждна система за морфологична анотация на средновековните старобългарски текстове, а нещата там не вървяха никак гладко. По искане на К. Симов и П. Осенова екип от трима души – А. Тотоманова, Т. Славова и Г. Ганева – разработи набор от тагове (тагсет) на всички възможни в старобългарски словоформи на базата на ***съществуващия тагсет за съвременния български език***, дефиниран и използван в синтактичния корпус BulTreeBank (215 149 думи)[3] . Съставеният от нас тагсет включва над 2000 тага и отчита различните фонетични варианти на завършеците на думите в средновековните славянски паметници, дължащи се на различните правописно-езикови редакции[4]. След тагсета разработихме и граматически речник на старобългарския език, който се състои от таблици, съдържащи всички словоизменителни парадигми на имената, местоименията и гаголите. И двата инструмента – тагсетът и граматическият речник, бяха представени на ***Лятната школа*** по проекта, проведена в края

[3] Описание на етикетите вж. у *Simov, Osenova, Slavcheva 2004*.

[4] Самият тагсет и принципите на неговото изготвяне са публиувани в сборника от заключителната конференция по проекта, вж. *Тотоманова, Славова, Ганева 2015*.

на м. юли – началото на август миналата година. Започна тяхното тестване върху текста на Бориловия синодик, който аз анотирах ръчно (с оглед на отделянето на думите и нормализацията[5]) няколко пъти. Резултатът от анотирането, базирано отново на CLaRK, беше още по-разочароващ от тестването на търсачката, защото, както беше докладвано на ***Лятната школа***, към август 2014 г. анотаторът можеше да разпознава само частите на речта, но не и останалите граматически категории. И тук отново ни се наложи да се обърнем към г-н Тодор Тодоров, който в момента тества прототип на морфосинтактичния анотатор върху граматическия речник на прилагателните, използвайки дигитализирания вариант на Старобългарския речник. Липсата на анотатор обаче се отрази неблагоприятно върху една от допълнителните задачи на проекта – изработването на речник-индекси и по-специално създаването на речник-индекс на Бориловия синодик и съпътстващите го в Палаузовия сборник чинопоследования и решения на православни събори. За изпълнението на тази задача проектът се възползва от помощта на доц. И. Христов, който е разработил специален софтуер за индексация на текстове и за подбор на гръцките съответствия, а морфосинтактичната анотация на среднобългарския текст (18 552 словоформи) трябваше да се извърши ръчно. Резултатите от работата по този речник-индекс бяха представени на поредния семинар по проекта, проведен на 17.06.2015 г., затова няма да ги повтарям, но съм сигурна, че новото издание ще бъде готово за представянето на финалния отчет по проекта в началото на септември[6].

Дейностите по попълването на диахронния корпус с нови текстове и по съставянето на дигиталната библиотека с електронни издания в областта на палеославистиката и медиевистиката, започнати по предишни проекти, обаче вървяха гладко и сега със задоволство можем да кажем, че почти удвоихме броя на текстовете в корпуса, които сега са над 130. По време на проекта в корпуса влязоха съчиненията на Климент Охридски, Йоан Екзарх, Константин Преславски, Патриарх Евтимий, Константин Костенечки; текстовете на Ефремовската кормчая, на Манасиевата хроника, на Диоптрата на Филип Монотроп, на Германовия, Тиквешкия и Берлинския сборник и други важни за нашата книжовна и езикова история паметници. Използвам случая, за да изкажа благодарността си към колегите Д. Пеев, М. Димитрова и останалите автори на ***Зографското издание на История славянобългарска***[7] и на издателите на ***Ловешкия дамаскин*** проф. Б. Велчева и О. Младенова[8], които позволиха текстовете им да бъдат включени в корпуса. Обновяването на софтуера на корпуса, за което споменах в на-

[5] Тук нормализацията означава свеждането на алографите до един-единствен знак.

[6] Вж. *Тотоманова, Христов 2015*.

[7] Вж. *История славянобългарска 2012*.

[8] Вж. *Велчева, Младенова 2013*.

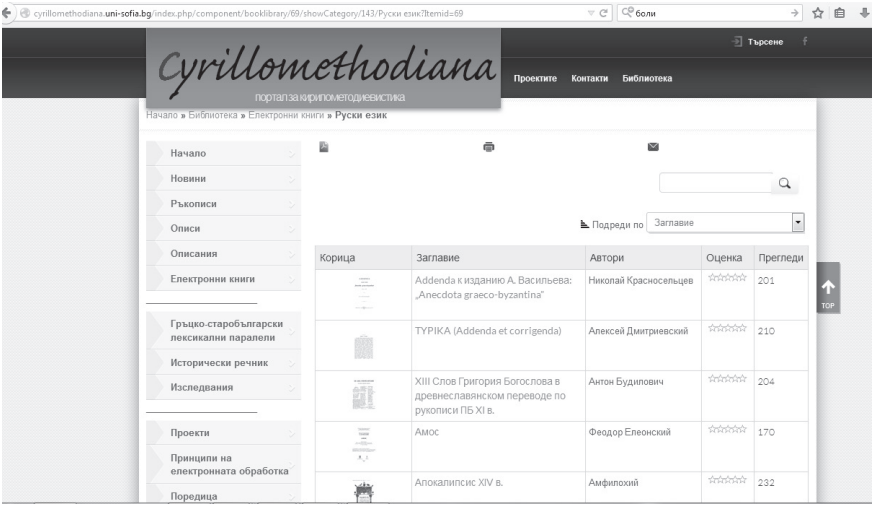
чалото, позволи в прозорчето с метаданните за публикуваните текстове да бъдат добавени нови функционалности, което направи възможно да се определя жанрът на публикуваните текстове чрез избора на съответната опция от падащото меню. За удобство на потребителите част от заглавията на текстовете бяха преведени и на латински.

Софтуерът на корпуса беше използван за дигиталното издание на Архивския хронограф, подготвено по проекта „История и историзъм в православния славянски свят. Изследване на идеите за история“, за което беше създаден и първият вариант на Unicode шрифтовете и конвертора. Пълният текст на Архивския хронограф може да се види редом с корпуса и речниците в рубриката Хронограф. Публикацията на тези текстове не само свидетелства за връзките на настоящия проект с други проектни инициативи, но и потвърждава виталността на проектната идея за използването на дигиталните технологии за изследването на историята на българския език и култура.

Фиг. 6. Електронно издание на Архивския хронограф

Електронната библиотека с издания на паметници, справочници и речници за нас е постоянен повод за задоволство. В момента тя съдържа над 600 заглавия на трудове по история, медиевистика, славистика и старобългаристика, които се намират на интернет на адрес http://cyrillomethodiana.uni-sofia.bg/index.php/component/booklibrary/218/all_category?Itemid=218, и е една от най-посещаваните рубрики на портала по кирилометодиевистика. Всички заглавия са снабдени с подробно изработено по всички правила на библиотечните науки описание на съответните издания и тези, които вече не подлежат на законите за авторско право, са на свободен достъп и могат да се копират от посетителите на страницата в ***pdf*** или ***DjVU*** формат. Библиотеката беше създадена с уси-

лията на колегите доц. Е. Мусакова и ас. д-р Поли Муканова, които въведоха и приложиха принципите за библиографска обработка към събраните от нас електронни издания.



Фиг. 7. Електронна библиотека

В заключение ще кажа, че въпреки трудностите проектът „Информатика, граматика, лексикография“ обогати нашите възможности за дигитална обработка на средновековните български текстове и ни даде нови идеи за използването на разработените от нас електронни инструменти. Така че интересното тепърва предстои...

ЦИТИРАНА ЛИТЕРАТУРА

Велчева, Младенова 2013: *Велчева, Б. и О. Младенова.* Ловешки дамаскин. Новобългарски паметник от XVII век. София, 2013, 536 с.

Илиева 2013: *Илиева, Т.* Терминологичната лексика в Йоан-Екзарховия превод „De fide orthodoxa“. София, 2013, 406 с.

История славянобългарска 2012: *Паисий Хилендарски,* „История Славянобългарска“, критическо издание с превод и коментар, издава Зографска света обител, Света гора, Атон, 2012 г., 413 с.

Тотоманова 2011: *Тотоманова, А.* Проектът „Компютърни и интерактивни средства за исторически езиковедски изследвания“ и дигиталното представяне на словното богатство на българския език през вековете. – В: Проект „Компютърни и интерактивни средства за исторически езико-

ведски изследвания“. Сборник доклади от заключителната конференция 15.12.2011. С. 2011, с. 5–15.

Тотоманова, Славова, Ганева 2015: *Тотоманова, А., Т. Славова, Г. Ганева.* Морфосинтактичен тагсет на старобългарския книжовен език. – В: Сборник доклади и материали от заключителната конференция. София, 29–30.06.2015 г.

Тотоманова, Христов 2015: *Тотоманова, А., И. Христов.* Речник-индекс на словоформите в Бориловия синодик и придружаващите го текстове в ръкопис НБКМ 289. София, 2015.

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

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

Totomanova 2012: *Totomanova, A.* Digital Presentation of Bulgarian Lexical Heritage. Towards an Electronic Dictionary. – В: Studia Ceranea 2, 2012, pp. 221–234.