МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Електронни инструменти за изследване на средновековната лексика и граматика

Анна-Мария Тотоманова, Гергана Ганева

Исторически речник, граматически речник, автоматичен морфологичен анотатор, електронни корпуси

Digital Tools for Studying Medieval Lexis and Grammar

Anna-Maria Totomanova, Gergana Ganeva

The present report summarizes the results of several projects focused on the development of digital tools for studying, adequate presentation and popularization of one part of Bulgaria's cultural and historical heritage: Bulgarian language and mediaeval literature. The system comprises an electronic diachronic corpus, a dictionary of Old Bulgarian, a historical dictionary, a grammatical dictionary, and an automatic morphological annotator.

Изграждането на уеб базирана система за представяне и изследване на средновековната лексика и граматика в диахронен план е трудоемка и дългосрочна задача, още повече че за пръв път се прави върху славянски материал. Затова в изграждането на системата са инвестирани усилията на екипите на три проекта: „История и историзъм в православния славянски свят. Изследване на идеите за история“. „Компютърни и интерактивни средства за исторически езиковедски изследвания“, „Информатика, граматика, лексикография“, като всеки един от проектите е със специфичен принос към поставената задача.

В идеалния случай подобна система трябва да включва следните инструменти: специализирани шрифтове по стандарта уникод, електронен диахронен корпус от дигитално набрани текстове, търсеща машина и виртуална клавиатура, граматически речник, електронни речници, анотатори (тагери).

Изграждането на системата (ние я нарекохме histdict) започна още през 2008 г. с разработването на специализирани старобългарски шрифтове с разнообразен инвентар от буквени и диакритични знаци, съобразени със стандарта уникод. Уникодските шрифтове се изграждат на базата на широко разпространените в България CyrillicaBulgarian, CyrillicaOhrid, CyrillicaShafarik и CyrillicaEpigraph, създадени от г-н Тодор Тодоров. Към момента разполагаме с два напълно окомплектовани средновековни шрифта CyrillicaBulgarian10U и CyrillicaOhrid10U и с един шрифт, предназначен за набиране на ранни новобългарски текстове, който е в процес на доразработка. Към двата шрифта беше разработен конвертор, който прехвърля текстове, набрани със старите шрифтове, към един от новите. Конверторът беше усъвършенстван на няколко пъти, като към него бяха добавени нови

опции. В последната си версия той конвертира и старобългарския PopRetkov шрифт, използван в Италия, и два старогръцки шрифта TmsGkOld, TmsGkClassic, които се конвертират към Palatino.

Началото на електронния диахронен корпус беше положено в края на 2011 г. и в него влязоха както конвертирани, така и новонабрани текстове. Корпусът беше замислен като основа за изработване на исторически електронно базиран речник на българския език. Целта на корпуса е да представи българското книжовно наследство от X до XVIII в. в цялото му жанрово и тематично многообразие. Насочваме се към текстове с доказано българско потекло — оригинални произведения или преводи на български книжовници. Съзнателно изоставихме (поне засега) текстовете от класическия старобългарски канон, защото те вече са обработени лексикографски. Документите в корпуса обикновено отразяват оригиналния правопис на паметниците или на изданията, от които са преписани. Корпусът е на свободен достъп и разполага с интрументи за текстова критика: коментари, разночетения, палеографски и археографски бележки. Към момента корпусът включва 110 документа.

Засега външните потребители на корпуса могат да търсят в отделните документи и по ключови думи от заглавията и археографските бележки. Членовете на екипа разполагат с търсеща машина, която позволява създаването на различни по тип конкорданси. Надяваме се в скоро време към нея да бъде изработен потребителски интерфейс, който да позволи използването ѝ от широк кръг потребители и за целите на всякакви изследвания. Системата histdict разполага и с виртуална клавиатура, която ще стане нейна интегрална част.

На платформата ще стоят два речника — дигитализираният Старобългарски речник в 2 т., изработен от ИБЕ-БАН (той впрочем отдавна е на разположение на всички потребители на сайта histdict.uni-sofia.bg), и електронно базираният исторически речник на българския език, за който беше разработен специализиран софтуер. Новият софтуерен продукт позволява както създаването на нови речникови статии онлайн, така и редактирането и обогатяването на вече съществуващите статии в старобългарския речник. Софтуерът позволява и всяка речникова статия да бъде снабдена с подробна граматическа информация. Затова в него се инкорпорира граматически речник на средновековния български книжовен език.

От половин година насам екипът на проекта разработва речникови статии за историческия речник. Лексикалните значения се извличат както от текстовете в корпуса, така и от съществуващите към момента лексикографски наръчници (речници, речник-индекси). Решихме първата група лексеми, които обработваме в диахронен план, да включва християнската терминология. За целта беше разработен словник, който включва близо 800 думи.

Свързването в една обща система на електронен диахронен корпус, софтуер за речникова статия и търсеща машина позволява да се изработват различни по тип лексикографски справочници: речници на отделни паметници (речник-индекси), речници на езика на даден книжовник, синхронни речници (напр. речник на

XIII в.), тематични речници, идеографски речници и др., като структурата на речниковата статия може да варира в зависимост от целта на лексикографското изследване. Например доц. Иван Христов изработи средновековен гръцко-български речник на базата на дигитализирания старобългарски речник (линк към този речник е сложен в системата histdict).

В хода на работа най-сложно се оказа изработването на автоматичен морфологичен анотатор (тагер), който машинно разпознава граматическите значения на думите в средновековните български текстове. Екипът на проекта вече е изработил детайлно описание на старобългарската граматика, което включва близо 3500 тага. Предизвикателствата при изработването на този инструмент произтичат от факта, че българският език е с богата флективна морфология и дълга писмена история. От друга страна, е необходимо тагерът да бъде свързан с останалите инструменти от системата histdict, където вече има и други инструменти за коментар към думите в текстовете.

Цялата система histdict е замислена като отворена и обогатяваща се платформа, което ще позволи нейното използване и усъвършенстване и от следващите поколения изследователи у нас и в чужбина.