Електронни инструменти за обработка на средновековни славянски текстове

От края на 2009 г. насам научен колектив, който включва членове на катедрата по кирилометодиевистика и на секцията по история на българския език в ИБЕ към БАН, колеги от други университети, както и специалисти по информатика, последователно и методично изработва набор от електронни инструменти за обработка на средновековни български и ранни новобългарски текстове. Финансирането на дейностите се осигурява от два последователни докторантски проекта по ОП „Човешки ресурси“. Първият от тях завърши през декември 2011 г. и получи високо признание от управляващия орган на програмата, след като беше обявен за най-успешния проект в схемата. Вторият проект започна през м. ноември миналата година и представлява своеобразно надграждане на предишния проект, като включва както изработване на нови инструменти, така и усъвършенстването на вече създадените електронни средства. И двата проекта преследват едни и същи цели:

1. Да въведем дигиталните технологии в палеославистиката и историческото славянско езикознание, като по този начин съкратим времето за обработка на средновековните славянски текстове.
2. Да привлечем към тази високо специализирана област на езикознанието повече млади хора, за които използването на ИКТ е част от естествената им среда.

Започнахме с изработката и усъвършенстването на няколко Уникод шрифта ***Cyrillica Bulgarian 10U*** и ***Cyrillica Ochrid10U***. В момента имаме и вариант на шрифт ***Cyrillica Old Style10U,*** предназначен за ранни новобългарски текстове, който обаче се нуждае от доработка. Шрифтовете са разработени на основата на комплекта старобългарски фонтове на фирмата „Синтезис Софт“ и съдържат богат набор от знаци, включително и диакритика, което позволява адекватното предаване на средновековните текстове.

Шрифтовете са вградени в конвертор, който има следните функции. (слайд). С помощта на конвертора успяхме да създадем основите на диахронния корпус на българския език, в който първоначално влязоха вече публикувани и/или набрани от членове на проектния колектив текстове. Материалите са набрани дигитално според преписа, в който са оцелели, или изданието, ако има такова, и благодарение на шрифта и развитието на технологиите са видими от всякакви браузъри.

Хронологичният обхват на корпуса е 10-19 век. Текстовете представят различни жанрове на старата ни литература – хроники, послания, богослужебни текстове, екзегетична литература, историко-апокалиптични съчинения, жития, похвални слова, дамаскини, монашеска литература, юридически текстове, въпросо-ответни съчинения и т.н.

Корпусът включва на първо място съчиненията и преводите на старобългарските писатели, както и неатрибуирани преводи и компилации с доказано българско потекло. Влизат и некнижовни писмени паметници – приписки, надписи и графити, грамоти и послания.

Базата данни е разположена на адрес ***http://histdict.uni-sofia.bg*** и в момента съдържа 109 различни текста, между които са: *Троянският дамаскин, Пандектите на Антиох, Отговорите на Псевдо-Кесарий, Бориловият синодик*. Някои от тях, като *Посланията на патриарх Фотий до българския цар Борис*, бяха публикувани за пръв път в корпуса преди хартиеното издание. За сравнение, към края на предходния проект – декември 2011 г. – документите в корпуса наброяваха 75. Текстовете, които са в процес на въвеждане и редактиране, не са видими за широката публика. Само членове на колектива с редакторски достъп имат право да правят промени във всички текстове, да ги публикуват на свободен достъп и да ги скриват за допълнително редактиране. Преписвачите на текстовете и тези, които ги въвеждат в системата,

имат достъп само до конкретните текстове, по които работят. Авторските права на публикуваните текстове се запазват.

Към момента са подготвени и въведени в корпуса съчиненията на Климент Охридски по направеното от Катедрата издание, но не са публикувани. Преписани са прозведенията на патриарх Евтимий по изданието на Калужняцки, Манасиевата хроника по Московското издание, Евтимиевият служебник. Компютърният набор на Азбучния патерик и на Римския патерик ни беше предоставен от техните италиански издатели с любезното съдействие на проф. Марио Капалдо. Те вече са конвертирани и също очакват своя ред да влязат в корпуса. Преписват се Сказанието за Енох, преводите на Йоан Екзарх по изданията на Айцетмюлер и Садник, Ефремовата кормчая по изданието на Бенешевич, Тиквешкият сборник. Колегата Д. Пеев ни предостави електронния вариант на новото издание на Паисиевата история. По договорка с колегите Вълчанови от Осло очакваме да получим и техния корпус от ранни руски преписи на старобългарски ръкописи[1], който техният университет вече няма интерес да поддържа.

Всеки от текстовете в корпуса се представя с необходимите метаданни: заглавие, паметник, сигнатура на използвания ръкопис, правопис, автор и т.н. В хода на работата установихме, че темплейтът се нуждае от допълнителна рубрика за евентуално посочване на изданието на текста. Проблем е и разделянето на старобългарския от съвременния кирилски шрифт в целия корпус. Тъй като броят на текстовете в корпуса бързо нараства (за сравнение корпусът на проекта ***Манускриптъ*** на Виктор Баранов съдържа около 90 средновековни текста[2]), предвиждаме в близко бъдеще да генерираме списък на заглавията на текстовете в корпуса, което ще улесни търсенето и използването на текстовете.

Софтуерът на корпуса включва възможности за коментирането и редактирането на текстовете с помощта на бутони за червенослов, за контекстуални бележки и разночетения. Тези функционалности превръщат корпуса в един чудесен инструмент за електронно издаване на средновековни славянски текстове. Същият софтуер ще бъде използван за електронното издание на Архивския хронограф, чийто текст по Московския препис вече е подготвен за електронно публикуване.

Действащата към момента търсачка позволява да се търсят текстове в корпуса по различни показатели – метаданни или наличието на отделни лексеми.

Диахронният текстови корпус е и чудесен инструмент за дигиталното представяне и популяризиране на Кирило-Методиевото книжовно наследство. Корпусът вече успешно се използва и за обучението на студенти, докторанти и специализанти, които имат интереси в областта на палеославистиката и историческата медиевистика и диахронното езикознание. Предстои свързването му и с разработваните по един друг проект, финансиран от ОП РЧР, *електронни ресурси за обучение по медиевистика*. Нещо повече, предвижда се обучителната платформа *e-medievalia* да бъде инсталирана в поддомейна *histdict.uni-sofia.bg*. За да се увеличи броят на потенциалните потребители на корпуса в бъдеще ще се включат и преводи на съдържащите се в него текстове на съвременен български и/или на английски език.

***

Към корпуса в момента се разработват и следните електронни инструменти:

---

1 Corpus of Old Slavic Texts from the XIth c. Introduction by R. Pavlova. Sofia–Trondhaim, 2001. [http://www.hf.ntnu.no/SofiaTrondheimCorpus/index2.html].

2 http://manuscripts.ru.

Търсеща машина, която да търси едновременно във всички текстове цели думи, по началата им, завършеците или отделни низове букви в тях. Търсачката е изключително необходима с оглед на основното предназначение на корпуса – създаване на електронно базиран исторически речник на българския език. Тя е нужна и с оглед на това, че почти всички текстове с изключение на нормализираните съчинения на св. Климент Охридски са представени в оригиналния им правопис – български, руски, сръбски и/или смесен, и една дума може да има различни графични варианти в зависимост от правописа и датировката на преписа. Търсачката ще бъда придружена с *виртуална клавиатура*, която ще улесни ползването й. В момента е възможно да се търсят отделни думи или низове от знаци само в рамките на отделно публикувания текст с редакторските бутони *Find* и *Replace*, познати ни от Word. Търсачката се разработва в близко сътрудничество с колегите П. Осенова и К. Симов, които ни предоставиха своя прототип за търсене в корпуси за съвременна българска реч по системата Clark. Остава да се изработи подходящ интерфейс и да се свърже с виртуалната клавиатура.

Морфологичен анотатор на средновековни старобългарски текстове. Към момента съществуват възможности за ръчно анотиране на текстовете от корпуса, като анотираният текст се запазва в отделно копие и може да бъде редактиран и публикуван според желанията на анотиращия. Граматичната информация се вписва в поредица от падащи менюта (темплейти), в които анотиращият избира необходимите показатели в замисимост от граматичната природа на думата. По текущия проект се разработва нов автоматичен анотатор. Първият етап от подготовката за изработването на този инструмент вече е завършен – трима от участниците в проекта подготвиха тагсет (набор от тагове/етикети) за морфологичен анализ на старобългарски текстове. Тагсетът описва граматиката на Кирило-Методиевия език и съдържа около две хиляди тага. В момента таговете се свързват със съответните форми (използва се Учебникът по старобългарси език за НГДЕК с автори И. Добрев, Ж. Икономова и А. Тотоманова) и се създава електронен лексикон. Предстои лексиконът да се свърже с формите в дигитализирания от предишния проект *Речник на старобългарския език*, изработен от ИБЕ при БАН, и анотаторът да се изпробва върху текстове от корпуса. След тестването предстои ръчно анотиране на автоматично анотирания текст и след това – писане на правила за снимане на омонимичността на формите. Софтуерът на анотатора е поверен на колегите П. Осенова и К. Симов, които имат опит с анотацията на текстове на различни езици.

Електронен инструмент за изработване на речник-индекси на отделни текстове и/или паметници. В момента разполагаме с един груб прототип, който е част от конвертора. Довършването му стана възможно едва сега, след като беше изработена и намерена дефиниция за азбуката (софтуерна) и установен буквеният ред. Съчетаването на двата инструмента – анотатора и индексатора, ще ни позволи бързата изработка на речник- индекси на различни текстове.

Създаденият диахронен корпус на българския език е съвместен с електронно базиран речник и е снабден със съответните инструменти за електронна обработка на текстовете. Електронните инструменти може да се използват:

1. **За създаването на електронно базирани лексикографски наръчници от всякакъв вид:**

- Диахронни исторически речници
- Исторически речници от синхронен тип (речници на книжнината на определен период, речници на езика на отделен книжовник или книжовно средище)[3]

3 Г. А. Богатова, цит. съч., с. 83-84.

- Речник-индекси на отделни паметници
- Тематични речници
- Етимологически (историко-етимологически) речници

2. **Исторически езиковедски изследвания в областта на:**
    - Морфологията и морфосинтаксиса
    - Морфонологията
    - Фонетиката и морфонологията
    - Лексикологията
    - Етимологията
    - Словообразуването
    - Фразеологията
    - Текстологията
    - Правописа

3. **За нуждите на университетското обучение на всички нива (бакалавър, магистър, доктор) в следните области:**
    - Старобългаристика и палеославистика
    - История на българския език
    - История на книжовния език
    - Стара българска литература
    - Средновековна история
    - Компютърна и корпусна лингвистика

4. **За подготвянето на електронни и хартиени издания на:**
    - Средновековни писмени паметници
    - Лексикографски наръчници
    - Учебни помагала (христоматии, речници, учебници)

5. **За популяризиране на българското културно и книжовно наследство у нас и в чужбина.**