Български език 57 (2010), 1, 7-22
ISSN 0005-4283

Андрей Бояджиев
Софийски университет „Св. Климент Охридски“
София, България

ЗА ЕЛЕКТРОННИЯ КОРПУС ОТ СРЕДНОВЕКОВНИ БЪЛГАРСКИ ТЕКСТОВЕ С РЕЧНИК

(Резюме)

Статията е преглед на бъдещата структура и технология за средновековния български корпус от текстове с речник. Съществуват едва няколко примера за исторически корпус на средновековни славянски езици изобщо: TITUS, Манускрипт и ССМН. Статията обединява няколко различни информационни източника: езиков корпус, електронен речник и различни структури от метаданни (описания на ръкописи, библиография, авторитетни файлове). В БАН и в българските университети съществуват много ръчно написани каталози на лексикални форми, които все още не са дигитализирани. Направен е опит да се опише възможния подход за представяне на речниковата форма и структура на електронния исторически словник и корпус на българския език с помощта на технологията на маркиращите езици. Основата на проекта използва няколко части от езика XML: TEI P5 модели за общата структура на текста, речника, библиографията и представянето на метаданните; моделът на инициативата Repertorium за описанието на средновековните славянски ръкописи. Модели са разширени с прибавянето на възможност за анотиране на частите на речта, като са проучени сегашните възможности на XML семейството от езици, като XSLT, XQuery и представянето на резултати в XML база от данни, например в Exist. Предложеният подход разчита на вече публикуваните стандарти, препоръки и техники за използване, които следват традициите, развити при обработката с компютърни средства на естествените езици.

Ключови думи: история на българския език, средновековен български език, езиков корпус, електронен речник, XML, XSLT, XQuery, Unicode, TEI

✉ Андрей Бояджиев,
aboy@slav.uni-sofia.bg

Публикувано: 31 март 2010