Савремени језички модели у
ɔоучавању српског културног наслеђа

Тршић 14-18.12.2026.

Универзитетска библиотека „Светозар Марковић“ (УБСМ) је централна академска библиотека Универзитета у Београду и матична високошколска библиотека на територији централне Србије. УБСМ у свом фонду поседује око 2 милиона јединица. Данас, УБСМ је савремени центар информационог друштва, друштвено препознати лидер у области дигитализације и институција са великим искуством у међународним пројектима. Једна од централних активности УБСМ јесте едукација библиотечке, студентске и истраживачке заједнице. УБСМ има широку скалу едукативних активности – од едукације истраживача о програмима, методама и ресурсима, преко едукације студената докторских студија, уз подршку Министарства науке, технолошког развоја и иновација, до едукација студената, библиотекара и истраживача из области дигиталне хуманистике, отворене науке, језичких технологија, примене вештачке интелигенције и других актуелних тема. УБСМ у редовима запослених има неколико међународно признатих експерата из ових области који су чланови међународних стручних тела, радних група, програмских одбора.

Друштво за језичке ресурсе и технологије (JePTex) је друштво основано ради остваривања циљева у области промоције, популаризације и унапређивања свих грана језичких технологија на научном, стручном и практичном нивоу. Чланови друштва су међународно препознати стручњаци, са огромним искуством у међународним пројектима, врхунским научним референцама и великим искуством у едукацији студентске и истраживачке заједнице. JePTex је пионир у креирању језичких технологија за српски језик и данас располаже највећим ресурсима, апликацијама и моделима за српски језик.

И УБСМ и JePTex су претходне три године успешно имплементирали петодневне радионице за студенте основних, мастер и докторских студија из области дигиталне хуманистике у Научно-образовном културном центру „Вук Караџић“ у Тршићу.

АУТОРИ ПРОГРАМА: Наведите ауторе програма додатног образовања и усавршавања талентованих ученика и студената за бављење научноистраживачким радом. Обавезно доставити радне биографије аутора програма.

др Василије Милновић, проф. др Ранка Станковић, проф. др Цветана Крстев, проф. др Душко Витас, др Михаило Шкорић

РЕАЛИЗАТОРИ ПРОГРАМА: прелиминарна листа предавача са њиховим звањима, афилијацијом (пун назив и седиште установе у којој је предавач запослен) и кратком биографијом.

Др Василије Милновић – Руководилац Центра за науку, Универзитетска библиотека „Светозар Марковић“, Булевар краља Александра 71, 11120 Београд. Руководилац неколико пројеката из области дигиталне хуманистике, српски представник у *European Open Science Cloud Steering Board*, Брисел и члан Радне групе LIBER (Association of European Research Libraries) *Digital Scholarship and Digital Cultural Heritage Collections Working Group*.

Проф. др Ранка Станковић – шеф Рачунарског центра Рударско-геолошког факултета (Ђушина 7, 11120 Београд) и професор на Катедри за примењену математику и информатику. Потпредседник је Друштва за језичке ресурсе и технологије JePTex. Руководилац је националног пројекта “TESLA - Text Embeddings - Serbian Language Applications” који финансира Фонд за науку Републике Србије, у оквиру ког се развијају разноврсни језички модели за српски језик. Главни и одговорни уредник часописа за дигиталну хуманистику „Инфотека“.

Проф. др Цветана Крстев – професор Филолошког факуктета Универзитета у Београду у пензији и суоснивач и председник Друштва за језичке ресурсе и технологије JePTex, координатор српског учешћа у COST Акцији *Distant Reading*, дугогодишњи уредник часописа за дигиталну хуманистику „Инфотека“ и аутор неколико електронских речника и корпуса српског језика.
Проф. др Душко Витас – оснивач и дугогодишњи шеф Катедре за рачунарство и информатику, у пензији, Математички факултет Универзитета у Београду, оснивач Друштва за језичке ресурсе и технологије и JePTex, саветник COST Акције *Distant Reading*, аутор СрпКор корпуса и неколико електронских речника и корпуса српског језика. Носилац ордена Републике Француске.
Др Александра Тртовац – библиотекар саветник Универзитетске библиотеке „Светозар Марковић“ (Булевар краља Александра 71, 11120 Београд), главни редактор каталога, члан Програмског одбора LIBER (Association of European Research Libraries) и члан Радне групе *Digital Scholarship and Digital Cultural Heritage Collections Working Group* исте асоцијације. Координатор неколико пројеката из области дигиталне хуманистике и главни и одговорни уредник часописа за дигиталну хуманистику „Инфотека“.
Др Михаило Шкорић – доктор интелигентних система, 8 година искуства у развоју алата у области језичких технологија, 5 година искуства у развоју језичких модела за српски језик; аутор апликација TExASe, PALMA, Дигитекс и бројних језичких модела, од претренираних до дообучених за конкретне задатке. Члан Јертеха, кључни истраживач на пројекту ТЕСЛА.

ПАРТНЕРИ У ОРГАНИЗАЦИЈИ ПРОГРАМА: Наведите установе/организације или друга правна или физичка лица која ће учествовати у организацији обука или као техничка подршка реализацији обука.

Друштво за језичке ресурсе и технологије - JePTex,
Ђушина 7, спрат 1, кабинет 152, 11120 Београд

КВАЛИТЕТ ПРОГРАМА ОБУКЕ: компетенције које програм развија, садржај,

НАЗИВ ПРОГРАМА: Савремени језички модели у проучавању српског културног наслеђа

ЦИЉ: Овај програм ће омогућити едукацију истраживачког кадра у контексту примене језичких технологија, која је ургентна и у најбољем националном интересу. Наиме, више није довољно само истраживати културно наслеђе, већ то треба чинити користећи савремене, технолошки напредне алате, како би само културно наслеђе, као и истраживање по себи имало већу видљивост и глобални карактер. Језичке технологије представљају интердисциплинарно поље које обједињује методе рачунарске лингвистике, обраде природног језика и вештачке интелигенције ради систематске дигиталне анализе текстуалних и језичких ресурса. Оне обухватају изградњу, обраду и анотацију корпуса, развој и примену језичких и мултимодалних модела, екстракцију и повезивање знања, као и интеграцију са дигитално-хуманистичким инфраструктурама, чиме се омогућавају нови истраживачки приступи проучавању српске културе и културног наслеђа. Самим коришћењем ових могућности, повећава се видљивост културе у међународној арени и обезбеђује савремени стручни интегритет истраживања. Како би процес био сасвим заокружен, неопходно је у њега укључити и библиотекаре, као највеће савезнике истраживача.

Посебан легитимитет овог семинара представља чињеница да се он наслања на активности које су успешно спроведене у оквиру пројеката *Удаљено читање* (2019), *Читање издалека* (2022), *Увод у дигиталну хуманистику* (2023), *Едукација библиотечке и истраживачке заједнице на моделу повећања видљивости српских историјских новина кроз напредне језичке моделе и технологије* (2024) и *Платформа за истраживање дигитализованог садржаја новинских колекција потпомогнута језичким технологијама за српски језик (ПИСАН)* подржаних од

стране Министарства културе Републике Србије, али и кровне европске акције у оквиру програма COST CA16204 - *Distant Reading for European Literary History*, која је започета 2017. године, а трајала до 2022. Такође, овогодишњи семинар ће бити заснован и на резултатима текућег пројекта *ТЕСЛА: векторизација језика – апликације за српски језик*, подржаног од стране Фонда за науку Р. Србије, као и текућег пројекта *SRB-NEWS-AI: Open Historical Serbian Newspaper Datasets for Low-Resource Language AI*, у оквиру *Microsoft* позива LINGUA. Битно је напоменути да ће се овај семинар наслањати и на претходна искуства рада у Тршићу 2023., 2024 и 2025. године.

ОБУХВАТ УЧЕСНИКА: 20 + талентованих студената основних, мастер и докторских студија филолошке и друштвено-хуманистичке оријентације, као и наставно особље изабраних стручних средњих школа и млади истраживачи.

ОЧЕКИВАНИ ИСХОДИ: оспособљавање полазника у припреми текстова за корпус/ дигиталну колекцију. Предвиђено је да радионице буду радног карактера. Полазници ће се кроз практичан рад упознати са целом линијом дигитализације: корекција сканираних и очитаних текстова, анотирање њихове структуре и изгледа, припремање мета-података. Напредним радионицама предвиђа се обучавање полазника за проверу и корекцију аутоматски обележених језичких објеката: имена људи, њихових улога, локација и друго. За ове потребе користиће се алати доступни на: http://ners.jerteh.rs/, https://portal.tesla.rgf.bg.ac.rs/, https://obrada.jerteh.rs/ и сл. Такође, полазници ће се упознати са применом језичких модела, повезивањем са отвореним подацима, електронским речницима, могућностима аутоматизације. Посебан сегмент ће се бавити проналажењем валидних ресурса и информација, као и дигиталних алата и њихова примена на текст. Кроз практичан рад, полазници ће усвојити веома много информација из области креирања језичких технологија, дигиталне хуманистике, али и удаљеног читања и отворене науке. Биће оспособљени да повећају видљивост свог рада, па самим тим и теме својих радова, а то је српска култура и језичко наслеђе у широком смислу.

ОБЛИЦИ РАДА: комбиновање теоријског и практичног рада. Осим теоријских излагања тема наведених у претходној секцији, полазници би имали прилику да припреме један текст, од корекција и обележавања TEI етикета до аутоматске анотације именованих ентитета. Припремљене текстове би полазници користили за повезивање са отвореним подацима, а потом би уследили практични задаци. Ресурси за практичан рад: srpELTeC корпус, колекција историјских новина - Претражива и језички модели развијени у оквиру пројекта ТЕСЛА.

ТРАЈАЊЕ АКТИВНОСТИ И ВРЕМЕНСКИ РАСПОРЕД АКТИВНОСТИ: предавања се организују 5 радних дана, пре подне, у периоду од 10-13 часова са краћим паузама, а радионице пет радних дана, после подне од 16-18 часова.

Преферирани термин – новембар-децембар 2026.

НАЧИН ПРАЋЕЊА ЕФЕКАТА АКТИВНОСТИ: свакодневно ангажовање полазника у практичном раду и излазна анкета полазника о семинару.

ДИЗАЈН АКТИВНОСТИ: Структура дизајна активности садржи: ток, ниво, главне поруке (кључне тачке), методе, аудиовизуелна средства, писани материјали, остали материјал и прибор.

Распоред за 5 дана:

Дан 1

Теоријски део (пре подне):

- Важност језичких технологија и дигиталне хуманистике данас, резултати досадашњих пројеката, циљеви примене језичких технологија у контексту културног наслеђа, језички ресурси и модели, дигитална хуманистика и отворена наука, питање канона и дигиталне хуманистике.

- процес дигитализације (скенирање, рашчитавање), алати који се користе. На примеру https://udaljenocitanje.unilib.rs/ и https://jerteh.github.io/PISAN/ демонстрација разлика слике и рашчитаног текста. Карактеристични проблеми за рашчитавање ћириличног и латиничног писма, проблеми старих и ретких књига, проблеми рашчитавања новина. Аутоматска корекција текста применом правила и језичких модела.

- Значај стандардизације (TEI, пример Вукових пословица, ELTeC-а, ПИСАН).

Практичан рад (после подне):

Сваки полазник креира сопствени мини пројекат у ком креће од рашчитавања скенираног документа (из колекције ПИСАН), након чега следе корекције текста и структурне анотације. Примери ће обухватити оба писма.

Дан 2

Теоријски део (пре подне):

- Обрада дигиталних текстова. Сегментација текста на реченице и токенизација. Врсте и значај анотација.

- Аутоматстка анотација токена врстама речи и лемама, аутоматска анотација делова текста именованим ентитетима (места, особе, организације, професије,...) у историјским текстовима. Изазови препознавања личности, места, институција.

- Текстометрија. Основна статистичка анализа корпуса. Претраживање и једноставна анализа садржаја. Партиције, подкорпуси и анализа специфичности.

- Напомене о напреднијим врстама анотације: синтакса, директан говор, семантика (значења), стилске фигуре. Неопходност контроле и евалуације.

Практичан рад (после подне):

Контрола ручне анотација структуре документа сагласно TEI препорукама (добро формирани и валидни XML документи); аутоматска анотација врстама речи и лемама (http://obrada.jerteh.rs/), именованим ентитетима (https://ners.jerteh.rs/). Контрола и евалуација аутоматске анотације (http://inception.jerteh.rs/).

Упознавање са корпусима доступним на https://noske.jerteh.rs/. Основе CQL језика и демонстрација упита над корпусом srpELTeC. Анализа резултата коришћењем конкорданци, фреквенција, извоз резултата. Инсталација алата TXM.

Дан 3

Теоријски део (пре подне):

- Отворени подаци, упознавање са википодацима о српским романима: основни подаци о роману и његовим издањима, подаци о аутору, о местима радње, главним ликовима и њиховим везама. Ручни и аутоматски унос википодатака.

- Демонстрација рада са Википодацима из пајтон свески.

- Претраживање података језиком SPARQL и визуелизација резултата табеларно, у виду мапе, графа, стабла. Интеграција резултата упита у презентације.

Практичан рад (после подне):

Анализа појављивања личности и топонима у новинама. Повезивање ентитета са спољним изворима знања. Визуелизација мрежа ентитета. Сваки полазник ће унети задате податке Википодатке, креирати упите са различитим излазима и интегрисати их са документом (HTML). Повезивање ентитета у тексту са Википодацима. Аутоматско препознавање именованих ентитета.

Дан 4

Теоријски део (пре подне):

- Улога библиотека у дигиталној хуманистици: пружање приступа дигиталним збиркама и базама података, обука корисника и подршка у коришћењу дигиталних технологија за анализу и обраду података из уманистичких наука; очување дигиталне културне и научне баштине (дигиталних копија старих рукописа и других важних историјских докумената). Конкретна улога библиотекара у дигиталној хуманистици на практичном примеру прикупљања материјала за припрему корпуса. Ресурс: историјске новине из колекције Претражива дигитална библиотека.

Практичан рад (после подне):

рад у програмима *Transkribus и Nanonets* (https://docstrange.nanonets.com/app/), напредно претраживање у COBISS-у. Историјске новине као истраживачки ресурс.

Дан 5

Теоријски део (пре подне):

- Читање у дигиталном окружењу - е-књиге. Текстометријска анализа текста коришћењем алата TXM: креирање корпуса, генерисање речника текста, конкорданце и фреквенције CQL образаца, креирање и анализа партиција и подкорпуса, анализа специфичности.

- Шири контекст српског језика у доба револуције вештачке интелигенције (ВИ), могућности и проблеми у вези са обрадом српског језика коришћењем ВИ, посебно у контексту великих језичких модела (попут OpenAI GPT), који тренутно привлаче глобалну пажњу. Статички и динамички језички модели. Иницијатива за развој јавно доступних квалитетних ресурса и алата за обраду српског коришћењем модела ВИ, како би се сачувала позиција српског језика у доба АИ револуције.

Практичан рад (после подне):

- Креирање сопственог малог корпуса текстова и његова текстометријска анализа.

- Коришћење модела за српски са платформе https://huggingface.co/jerteh і https://huggingface.co/te-sla.

- Инструкције (промптни инжењеринг) на примеру ChatGPT и других модела.

Семинар подразумева личне лап топ уређаје полазника, PPT презентације, пројектор, интернет, дигиталне колекције.