Нове технологије за оживљавање старих текстова

Цветана Крстев, Ранка Станковић, Бранислава Шандрих Тодоровић, Милица Иконић Нешић

ДИГИТАЛНА ХУМАНИСТИКА И СЛОВЕНСКО КУЛТУРНО НАСЛЕЂЕ II

– Зборник радова –

ДИГИТАЛНА ХУМАНИСТИКА И СЛОВЕНСКО КУЛТУРНО НАСЛЕЂЕ II

- Зборник радова –

Међународна научна конференција

Београд, 28. и 29. јуни 2021.

Уреднице:
проф. др Александра Вранеш
проф. др Љиљана Бајић
проф. др Љиљана Марковић

Савез славистичких друштава Србије
Београд, 2023

ДИГИТАЛНА ХУМАНИСТИКА И СЛОВЕНСКО КУЛТУРНО НАСЛЕЂЕ II

Друга међународна научна конференција Дигитална хуманистика и словенско културно наслеђе одржана је 28. и 29. јуна 2021. године, у организацији Комисије за електронске библиотеке и дигиталну хуманистику, Комисије за словенски свет и културе на Путу свиле Међународног комитета слависта (МКС), које су акредитоване током XVI Међународног конгреса слависта у Београду 2018. и Комисије за наставу словенских језика и књижевности, која је акредитована на XIV Међународном конгресу слависта у Охриду 2008. године.

САДРЖАЈ

Цветана Крстев
Универзитет у Београду
Филолошки факултет
cvetana@matf.bg.ac.rs

Ранка Станковић
Универзитет у Београду
Рударско-геолошки факултет
ranka.stankovic@rgf.bg.ac.rs

Бранислава Шандрих Тодоровић
Универзитет у Београду
Филолошки факултет
branislava.sandrih@fil.bg.ac.rs

Милица Иконић Нешић
Универзитет у Београду
Филолошки факултет
milica.ikonic.nesic@fil.bg.ac.rs

НОВЕ ТЕХНОЛОГИЈЕ ЗА ОЖИВЉАВАЊЕ СТАРИХ ТЕКСТОВА

Удаљено читање је парадигма која подразумева коришћење рачунарских метода за анализу великих колекција књижевних текстова. Да би се методе удаљеног читања могле применити, потребна је пажљива селекција текстова по одабраним критеријумима и њихова припрема. Управо овим се бави COST акција „Удаљено читање за европску историју књижевности“. Један од најважнијих циљева ове акције је припрема вишејезичног, прецизно балансираног корпуса који ће, када буде потпуно завршен, садржати по 100 романа први пут објављених у периоду 1840-1920. године за више европских језика, укључујући и српски.

Кључне речи: удаљено читање, књижевни корпус, обрада српског језика, анотација врстом речи, лематизација, именовани ентитети

1. Удаљено читање (D-READING) и корпус ELTeC

Удаљено читање је парадигма која подразумева коришћење рачунарских метода за анализу великих колекција књижевних текстова који обично потичу из дигиталних библиотека. Циљ ових анализа је допуњавање метода које се користе у теорији и историји књижевности. Термин „distant reading" се приписује Франку Моретију (Franco Moretti) и његовом раду „Conjectures on World Literature" (Moretti 2000). У овом раду Морети предлаже методе читања које укључују дела изван утврђеног књижевног канона који назива „the great unread". Новина коју за изучавање књижевности Морети предлаже је коришћење узорака, статистике, паратекстова и друга својства која се до тада нису обично користила у изучавању књижевности.

Да би се методе удаљеног читања могле применити, потребна је пажљива селекција текстова по одабраним критеријумима и њихова припрема. Управо овим се бави COST акција CA16204 *Distant Reading for European Literary History* (Удаљено читање за европску историју књижевности)[1] на којој је рад отпочео 2017. године, а завршава се почетком 2022. Један од најважнијих циљева ове акције је припрема вишејезичног корпуса (названог European Literary Text Collection - ELTeC) који ће, када буде потпуно завршен, садржати по 100 романа први пут објављених у периоду 1840-1920. за велики број европских језика, који чине језичке подколекције (Odebrecht et al. 2021).

Израда једног овако амбициозно замишљеног вишејезичног корпуса захтевала је брижљиву припрему. Тако су прво прецизирани критеријуми одабира дела за корпус који морају бити задовољени да би неко дело било уврштено:

- У обзир долазе само романи, то јест наративна проза (роман, новела или дужа приповетка) чија је дужина најмање 10.000 речи,[2] што значи да не долазе у обзир дела као што су путописи, есеји, биографије, аутобиографије, историјски списи и слично.
- Прво издање дела треба буде из периода од 1840. до 1920, укључујући и ове године.

---

[1] https://www.distant-reading.net

[2] Подразумева се да се речи броје аутоматски, на пример онако како то уради програм MS Word.

- Дело треба да буде оригинално написано на језику у чију ће се подколекцију уврстити, што значи да се преводи не узимају у обзир.
- Дело треба да буде објављено у Европи највише десет година после првог издања. Ова се одредба односи пре свега на дела на, рецимо, енглеском или португалском језику која су први пут могла бити објављена у Америци односно Бразилу.
- Предност се даје делима која су у назначеном периоду објављена као књиге, а не у наставцима у серијским публикацијама.

Осим ових обавезних критеријума, постављени су и додатни услови за састав сваке подколекције који треба, с једне стране, да обезбеде разноврсност заступљених текстова и, с друге стране, да омогуће компаративну анализу подколекција и примену кључних метода за статистичку анализу текстова. Ови додатни критеријуми за пожељну балансираност корпуса су следећи:

- *Величина колекције*: подколекција би требало да садржи 100 дела која се квалификују као романи (према претходно поменутим обавезним критеријумима).
- *Пол аутора*: подколекција би требало да садржи дела која су писали и мушки и женски аутори, а пожељно је да су 30% одабраних дела написале жене, а најмање 10%.
- *Број поновљених издања*: подколекција би требало да садржи како дела из канона, дакле добро позната широј публици, тако и потпуно непозната и заборављена дела. Одлучено је да се број поновљених издања неког дела узме као мера његове каноничности, па првој категорији припадају сва дела која су у периоду 1970-2010. имала бар два издања, док сва остала припадају другој категорији. Ових других би требало да буде бар 30%, али не више од 70%.
- *Равномерна покривеност периода 1840-1920*: Одабрани временски период првог издања дела је подељен у 4 периода у трајању од 20 година (само последњи период покрива 21 годину). Ови временски периоди би требало да буду равномерно заступљени у свакој подколекцији и да сваком од њих припада 20-25 дела.
- *Дужина дела*: Дела се према својој дужини деле у кратка (она која имају 10.000-50.000 речи), средње дужине (50.001-100.000) и дугачка (имају више од 100.000 речи). Подколекција би требало да садржи бар 20% дела свих дужина, а идеално би било 30-40%.

- *Број романа по аутору*: Подколекција би требало да садржи 9 до 11 аутора који ће бити заступљени са тачно 3 дела (што би, рецимо, омогућило тестирање аутоматских система за проверу ауторства), док би сва остала дела требало да буду написана од стране различитих аутора да би се обезбедила разноврсност. Уколико је за неке колекције тешко да се задовоље остали поменути критеријуми балансираности, ограничен број аутора може да буде заступљен и са два дела.

У овом тренутку[3] се припремају колекције за италијански, норвешки, румунски, украјински, хрватски, швајцарски, немачки, шведски, шпански, док су подколекције за енглески, мађарски, немачки, пољски, португалски, словеначки, српски, француски и чешки попуњене са 100 дела.

2. Подколекција српских романа SrpELTeC

Израдом подколекције српских романа (названа SrpELTeC) бави се тим истраживача, аутора овог рада, којим руководи проф. др Цветана Крстев. На основу увида у постављене критеријуме одабира и добре балансираности, представљене у претходном одељку, јасно је да израда подколекције српских романа није тривијалан задатак, и да је за његово извршење, што је случај и са многим другим европским језицима, потребно много више труда него, рецимо, за енглески или француски језик. Пре свега, прозно стварање на српском језику јавља се касније него у другим европским земљама, с појавом реализма који као правац преовлађује у последње три деценије XIX века (Деретић 1983:362), што практично значи да скуп из кога се дела могу бирати ради што бољег задовољења постављених критеријума за српски није тако богат као за друге језике који су у овом периоду имали развијенију књижевност. С друге стране, већина књижевних дела из периода 1840-1920, с обзиром да више не подлежу зашити ауторских права, дигитализована је за енглески, француски и неке друге европске језике и слободно доступна кроз дигиталне библиотеке,[4] што за српски језик није случај. Прецизније

---

[3] Стање ELTeC корпуса на дан 16. септембра 2021. године (ELTeC summary (distantreading.github.io)

[4] На пример за енглески језик коришћена је, између осталог, библиотека електронских књига Гутенберг (https://www.gutenberg.org/), за француски Викиизворник и

говорећи, многа дела српске књижевности из овог периода су дигитализована, али то није урађено на начин да би се она могла употребити за израду српске SrpELTeC подколекције. Најбогатија и најчешће коришћена таква библиотека је Антологија српске књижевности - АСК[5] урађена на Учитељском факултету Универзитета у Београду у сарадњи са Microsoft-ом. На жалост, овим дигиталним издањима недостају метаподаци тако да није познато која издања дела су коришћена за дигитализацију.

Због свега овога одлучено је да се српска подколекција изради у потпуности из почетка. Израда SrpELTeC подколекције одвијала се у више корака који ће бити укратко описани у наредним одељцима.

2.1. Одабир и проналажење српских 100 романа

Први, а можда и најтежи задатак, било је сачињавање листе дела која задовољавају критеријуме одабира. Велику помоћ у обављању овог изазовног задатка пружили су др Александра Тртовац и др Василије Милновић из Универзитетске библиотеке „Светозар Марковић“, проф. др Душко Витас из Друштва за језичке технологије и ресурсе. За почетно попуњавање ове листе послужиле су информације из дела „Српски роман 1800-1950“ (Деретић 1981) и Приповедачи (Милисавац 1978). Међутим, како се у овим књигама говори углавном о значајним романописцима и другим приповедачима и њиховим делима, листа је била непотпуна и готово да није садржала „маргинална“ дела српске књижевности, она која су објављена само једном или два пута и чији су аутори данас углавном непознати. Листа је даље допуњавана делима која су пронађена претраживањем заједничког каталога српских библиотека COBISS+ одабиром одговарајућих вредности за врсту садржаја или литерарни жанр (роман, приповетка, кратка проза и сл.). Понеки интересантни наслови пронађени су код београдских антиквара, а неке сугестије су добијене из већ прибављених старих књига које су на крају садржале листу објављених дела истог издавача.

После формирања ове листе прешло се на други корак, а то је проналажење самих књига и њихово сканирање. Одлука је донета да се скенирају прва издања кад год је то могуће, тј. када се прво издање може пронаћи у некој од библиотека са којима је остварена сарадња. Такође, ако је дело

Дигитална библиотека Квебека (Bibliothèque électronique du Québec), за мађарски Аустријска национална библиотека и Асоцијација мађараских дигиталних библиотека итд.

[5] http://www.antologijasrpskeknjizevnosti.rs/

првобитно изашло у наставцима у серијској публикацији, сканирано је његово прво издање у форми књиге, док се као година првог издања води издање из серијске публикације. Понекада се морало одустати од првог издања, јер је копија била лошег квалитета па је даљи рад био немогућ због изузетно лошег квалитета оптичког читања карактера. Такав је био случај, на пример, са романом „Општинско дете“ Бранислава Нушића, чије се прво издање из 1902. године није могло употребити, па је за даљи рад узето издање из 1932. године. Од неких дела из прве половине XIX века се морало одустати јер су писана предвуковском азбуком, а касније нису прилагођена, као на пример „Венацъ искренне любови Светоміра и Зорице : романтическа повѣсть сочинѣна Димитріемъ Михаиловићемъер“ из 1840. године. Ова дела се не могу обрађивати алатима за обраду савременог српског језика, па она из тога разлога нису уврштена.

На овај начин смо успели да сачинимо листу од више од 150 дела која би се потенцијално могла укључити у подколекцију SrpELTeC. Треба имати у виду да се пре прегледања самих књига и њихове дигитализације не може са сигурношћу рећи да ли се неко дело квалификује као роман, новела или дужа приповетка, па самим тим задовољава критеријум одабира, нити се може увек знати унапред да ли дело има више од 10.000 речи.

2.2. Дигитализација

Сканирање дела одабраних за SrpELTeC је обављено у библиотекама са којима је остварена сарадња и која су имале тражене примерке:

1. Универзитетска библиотека „Светозар Марковић“ (највећи део);
2. Народна библиотека Србије;
3. Библиотека Матице српске;
4. Библиотека САНУ;
5. Библиотека Катедре за Српску књижевност Филолошког факултета;
6. Приватна библиотека Душка Витаса и Цветане Крстев.

На крају су ипак коришћена и три дигитална издања из библиотеке ACK, а једини разлог је био убрзање рада на српској подколекцији.

После сканирања је уследила даља обрада која се одвијала у неколико корака:

1. Урађено је оптичко читање карактера (OCR) да би се из слике добио текст.[6]

2. Квалитет текста добијеног ишчитавањем карактера је варирао од текста до текста. У неким, ређим случајевима је био доста добар, у некима прихватљив, а у неким, такође ређим случајевима, толико лош да се није могао користити. Сви неодбачени текстови су се у сваком случају морали кориговати, што је прво рађено аутоматски, како је описано у (Krstev and Stanković 2019). Укратко, систем за аутоматску корекцију пореди све речи из текста са електронским морфолошким речником српског језика (Krstev 2008), све неупарене речи третира као погрешне и замењује их низом речи из истог речника које су могле произвести погрешно ишчитану реч као резултат уобичајених грешака ишчитавања. Ове уобичајене грешке нису увек исте, па се систем мора прилагодити сваком конкретном тексту. Једна од најчешћих грешака ишчитавања карактера у ћириличном тексту је замена „и“ са „п“ (и обрнуто), „п“ са „н“ (и обрнуто) и „н“ са „и“ (и обрнуто). Низ потенцијалних замена погрешних речи може бити празан (што значи да је реч можда и тачно ишчитана, али није забележена у речнику), садржати једног или више кандидата.

3. У великом броју случајева, ишчитани текст је тек после аутоматске корекције описане у претходној тачки био у таквом облику да се уопште може читати. Сваки овако коригован текст је потом читао читач-волонтер („акцијаш“), који је поредио текст с оригиналом, кориговао преостале грешке и бирао правог кандидата тамо где је било понуђено више. Акцијаши су анимирани преко Друштва за језичке ресурсе и технологије JeРТех,[7] њих укупно 25. Упутство читачима је било да текст треба да остане веран оригиналу, то јест да не треба да исправљају грешке из штампане верзије, а посебно да текст не прилагођавају савременом правопису (спојено или растављено писање, употреба малог или великог слова и слично).

4. Трећа и последња контрола састојала се од поновног поређења текста са електронским речником српског; непрепознате речи су представљале или преостале грешке које су исправљене, или речи које су недостајале у електронском речнику па је речник њима обогаћен, или су

---

[6] За овај задатак је коришћен ABBY FineReader, приватна верзија Д. Витас и Ц. Крстев.

[7] http://jerteh.rs/

представљале неку специфичност текста у погледу правописа или лексике што је остављано непромењено.

2.3. Анотација текста

Рад на корпусу ELTeC предвиђа за све подколекције основну анотацију, такозвану анотацију нивоа 1. Анотација се састоји од обележавања основних структурних елемената текста (поглавља и друге целине) и неких основних текстуалних елемената. Анотација је урађена у складу са TEI препорукама (TEI Consortium 2021), при чему је из богатог скупа елемената које ове препоруке дефинишу одабран као обавезан или дозвољен само мали подскуп.

Структурални елементи су:

5. Најосновнији је елемент који ће се наћи на почетку сваког поглавља, а ако је роман подељен и у веће или мање целине, онда се свако од њих означава са .
6. Ако се у роману појављује нешто што наликује песми – у виду засебних група редова – користе се обележја <quote> за целу „песму“ и <l> (line) за појединачне редове. Елемент <quote> може да се користи и за друге наводе (нпр. за цитирање делова неког другог текста, епиграфе на почетку целог текста или поглавља).
7. За означавање почетка нове странице у штампаном делу користи се ознака, нпр. <pb n="55"/>. Ове ознаке су врло корисне за кориговање текста, а исто тако и за упоредно приказивање скенираног и ишчитаног текста, какво је остварено у дигиталној библиотеци „Удаљено читање“ Универзитетске библиотеке „Светозар Марковић“.[8]
8. Подела у пасусе који се означавају етикетама <p> је обавезна, а за SrpELTeC су етикете додате аутоматски, на основу тврдог краја реда који оптичко читање карактера углавном задржава, а читачи проверавају у току кориговања текста.

Дозвољени су следећи текстуални елементи:

---

[8] https://udaljenocitanje.unilib.rs/

9. Ако се у тексту помиње неки наслов (најчешће је дат у курзиву или под наводницима) – наслов новина, књиге, позоришне представе и сл. – он се обележава етикетом <title>;
10. Ако се у тексту појављује део на страном језику (и он може бити у курзиву, али не мора), обележава се етикетом <foreign> којој се мора додати атрибут језика, нпр. <foreign xml:lang="FR">;
11. Део текста који је некако истакнут (курзивом, масним словима, подвучено, повећавањем величине слова, итд.), а није ништа од претходног, означава се етикетом <hi> (highlighted).

За сваки текст анотиран у складу с TEI препорукама обавезно је заглавље с метаподацима, па је то случај и са свим текстовима из корпуса ELTeC. На новоу акције је договорено који елементи заглавља су обавезни, тако да су заглавља свих колекција уједначена. Сва заглавља морају да садрже:

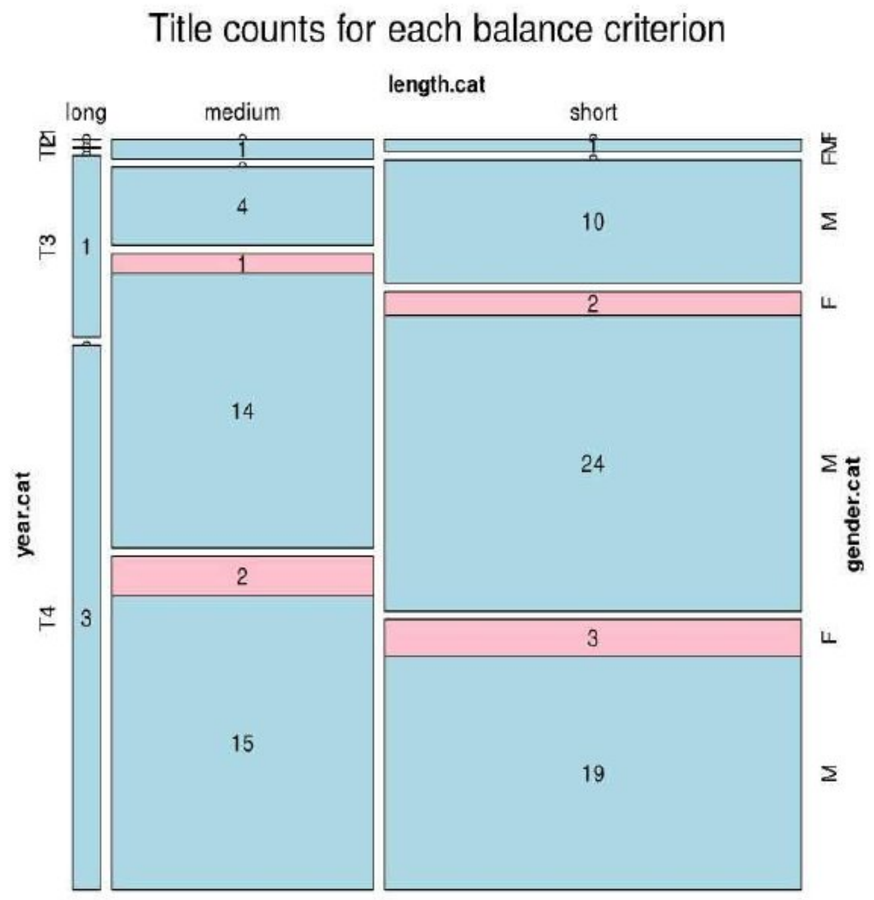
1. Опис електронског издања, што укључује назив дела и име аутора као и изјаве о одговорности (сканирање, корекција, анотација), датум публиковања, величина (мерена бројем речи). Аутору и делу могу се придружити идентификатори, као што су VIAF[9] и Wikidata.[10]
2. Кратак каталошки опис првог издања и издања које је коришћено за извор за ELTeC (уколико се разликује од првог издања).
3. Опис текста у смислу задовољавања критеријума балансираности.
4. Преглед свих измена дигиталног издања од његовог првог публиковања.

[9] VIAF: The Virtual International Authority File (http://viaf.org/)

[10] Слободна база знања (https://www.wikidata.org/wiki/Wikidata:Main_Page)

3. Репрезентативност српског корпуса

On 2021-09-16 ELTeC-srp contains 100 texts containing 4674766 words



Слика 1. Стање SrpELTeC-а 16. септембра 2021.

Напредак у изради корпуса ELTeC се редовно прати, и периодично се ажурирају сумарни подаци за све подколекције.[11] Стање сваке од 17 подколекција које су тренутно у изради или завршене приказује се и визуелно, графиконом који је дат на слици 1. Из заглавља графикона види се да је дана 16. септембра 2021. године српска подколекција садржала 100 романа укупне дужине од нешто више од 4,5 милиона речи (Odebrecht et al. 2021). Сам графикон приказује заступљеност романа по три критеријума: пол аутора, дужина дела и период у коме је дело први пут објављено. Пол аутора је представљен бојом: плавом бојом је означени мушки, а ружичастом женски. Одмах се уочава да су жене као аутори недовољно заступљене - само 8 од 100 дела. Дужина дела је представљена вертикалном поделом, и то с лева на десно: дугачка, средња и кратка. Уочава се веома мала заступљеност „дугачких“ романа (само 4), док преовлађују „кратки“ романи. Временски период је представљен хоризонталном поделом, одозго наниже иду први период (T1=[1840-1859]) до четвртог периода (T4=[1900-1920]). И

[11] https://distantreading.github.io/ELTeC/index.html

овде су уочава да је први период веома слабо заступљен са свега два дела, други нешто боље (14), док су преостала два периода равномерно заступљена са по 42 романа.

Ни мало не чуди да су неке од завршених подколекција учиниле много бољи избор романа те су њихове колекције готово савршено балансиране. Узмимо за пример француску подколекцију у којој је заступљено 34 жена-аутора, 30 дугачких романа, по 25 романа у сваком од четири временска периода, 44 често објављивана романа према 56 ретко објављиваних дела, а сличан је случај и са енглеском, немачком, португалском и мађарском подколекцијом. С друге стране, словеначка подколекција показује слична одступања од добре балансираности као и српска: 11 дела које су написале жене, 8 дугачких дела, 2 из првог периода T1.

Балансираност подколекција зависи од много фактора, од којих је један аутор подколекције: који људски и други ресурси су му на располагању.[12] Постоје и објективни фактори, а то је развој романа у назначеном периоду, о чему је већ било речи. Постоје и други „објективни“ фактори, а то је дужина романа мерена формално, бројем речи. Неки од романа из српске колекције, на пример „Ђурађ Бранковић“ Јакова Игњатовића, када би били преведени на енглески или француски, ушли би у категорију дугачких романа - у питању су карактеристике језика. Српска подколекција још није сасвим завршена, наиме неколико романа ће још бити замењено да би се постигла бар мало боља балансираност.

4. Даља обрада романа

Акција удаљеног читања предвиђа да све или већина подколекција буде анотирана и на нивоу 2, што подразумева детаљану анотацију: обележавање свих речи у тексту врстом речи[13] и лематизацију (додељивање свакој речи њеног уобичајеног речничког облика), опционо придруживање и других морфосинтаксичких описа, као и обележавање именованих ентитета.

---

[12] Треба имати у виду да се овај огромни подухват одвија у оквиру једне COST акције које финансирају само сарадњу, а не и конкретан рад, па је највећи део посла урађен волонтерски.

[13] https://universaldependencies.org/u/pos/

Главни циљ препознавања именованих ентитета (енг. *Named Entity Recognition, NER*) представља означавање имена особа, њихових улога, локалитета, организација у тексту, као и препознавање нумеричких израза који укључују датум, време, проценат, новац и представљају један од кључних корака при екстракцији информација из текста. Систем за препознавање именованих ентитета за српски језик је заснован на ручно креираним правилима која се ослањају на свеобухватне лексичке ресурсе за српски језик (Krstev et al. 2014).

На нивоу целе акције је договорено да се у романима означава само 7 категорија ентитета: PERS, ROLE, DEMO, ORG, LOC, WORK, EVENT, за које је закључено да ће бити од највећег значаја за даља литерарна изучавања (Frontini et al. 2020). Ове категорије су прецизиране на следећи начин:

PERS – укључена су властита имена људи: лично име, презиме, надимак или све то у комбинацији. Односи се и на имена стварних људи (*Марко Краљевић*), ликова из романа (*Чедомир Илић*), измишљених бића, што укључује и богове и свеце (*Зевс*). У оквиру имена су и додаци имену (као *Елизабета II*), док почасне титуле краљева, свештених лица и обичних људи (*краљ*, *др*, *г-ђа*) нису део имена већ се означавају са ROLE. У ову категорију спадају и имена животиња, ако су добиле властита имена. Присвојни придеви од личних имена се не обележавају.

ROLE – обележавају се занимања, титуле и задужења људи било да прате лично име особе или не. И овај ентитет се може састојати од више речи, нпр. генерал-пуковник, гимназијски професор. Треба водити рачуна о двозначности, нпр. „сељак“ може некада да означава карактерну особину, а некада занимање – само ово друго се обележава. Неке речи су двозначне, на пример, „бабица“ је деминутив речи „баба“, али може да означава и занимање (примаља).

ORG – означавају се имена компанија, политичких партија, образовних установа, спортских тимова, болница, музеја, библиотека, хотела, кафана, цркава и светилишта, на пример *Велика Школа*.

LOC – континенти, државе, региони, насељена места (градови и села), планине (острва, пећине, равнице...), водене површине (реке, океани, мора, језера, извори...), имена небеских тела, градске локације (имена улица, тргова, делова града). Као и у случају имена људи,

обележавају се и стварне и измишљене локације, па и онда ако су у тексту дате само иницијалом (нпр. дошао је из *Н.*).

WORK – наслови књига, драма, песама (за читање или певање), музичких дела, слика, скулптура, новина, на пример *Мали журнал*.

EVENT – имена догађаја који се редовно понављају или су се једном догодили али имају своје име, као што су природне катастрофе, револуције, битке, ратови, демонстрације, концерти, спортски догађаји, на пример, *Косовски бој* и *Митровдан*.

DEMO – обележавају се становници држава, градова, региона (*Турци*, *Турска*), или етничке групе (*Власи*), као и придеви настали од ових назива (*турски*).

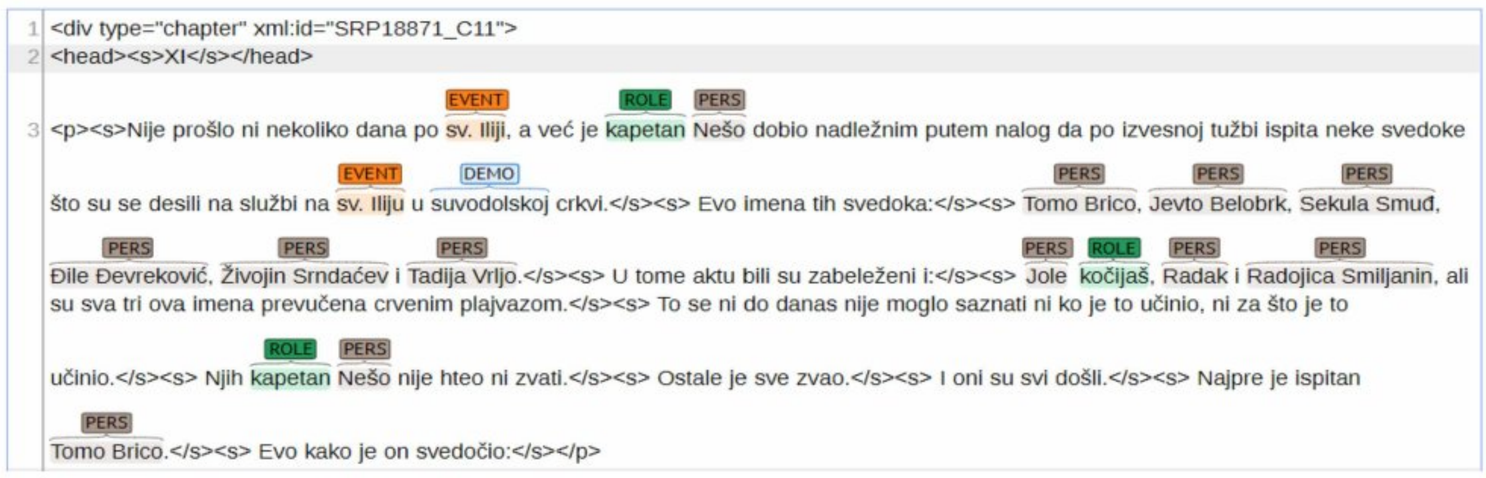
Сам процес анотирања романа одвија се за српски језик у неколико фаза. Прва фаза је аутоматско анотирање коришћењем SrpNER система за препознавање именованих ентитета (Krstev et al. 2014), који је првенствено намењен препознавању именованих ентитета новинских чланака. С обзиром на ову његову првобитну намену, систем се подешава за сваки роман понаособ ради постизања бољих резултата. То је могуће с обзиром да је у питању систем занован на правилима, те се ова правила могу мењати према потреби. У другој фази, произведене етикете, које садрже много више категорија и подкатегорија од 7 предвиђених ELTeC корпусом, аутоматски се прилагођавају скупу етикета које су у складу са ELTeC колекцијом. Трећа фаза је ручна евалуација у којој је укључено више акцијаша.[14] Евалуатори обављају корекције коришћењем два алата за анотацију Brat[15] и INCEpTION (види слику 2).[16]

---

[14] Евалуатори су студенти Универзитета у Београду разних нивоа: основних студија Библиотекарства и информатике на Филолошком факултету, мастер студија Рачунарство у друштвеним наукама и докторских студија Интелигентни системи.

[15] BRAT, https://brat.nlplab.org

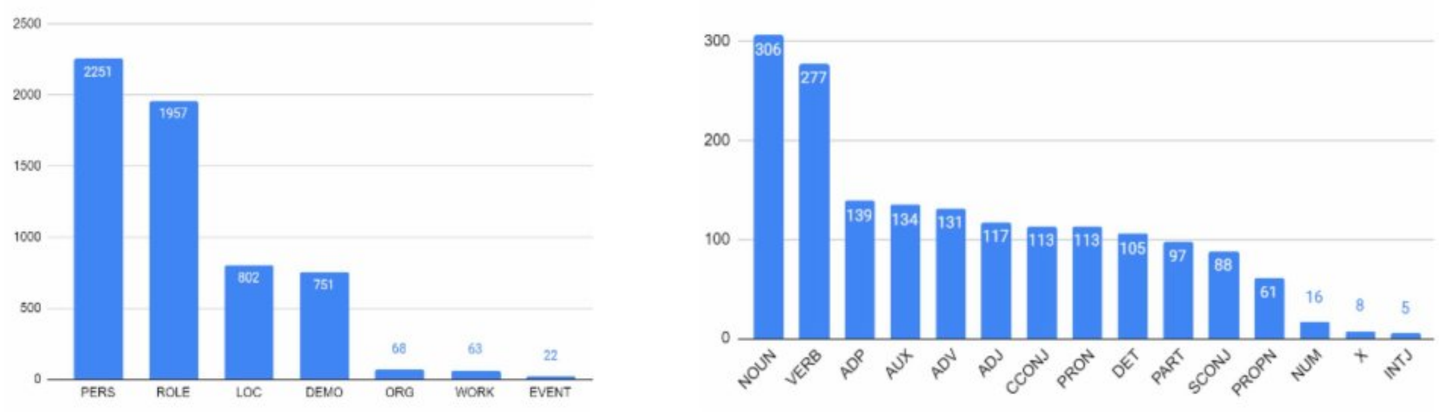
[16] INCEpTION, https://inception-project.github.io/



Слика 2. Именовани ентитети у роману Лазара Комарчића „Мој кочијаш“ визуелизовани кроз INCEpTION.

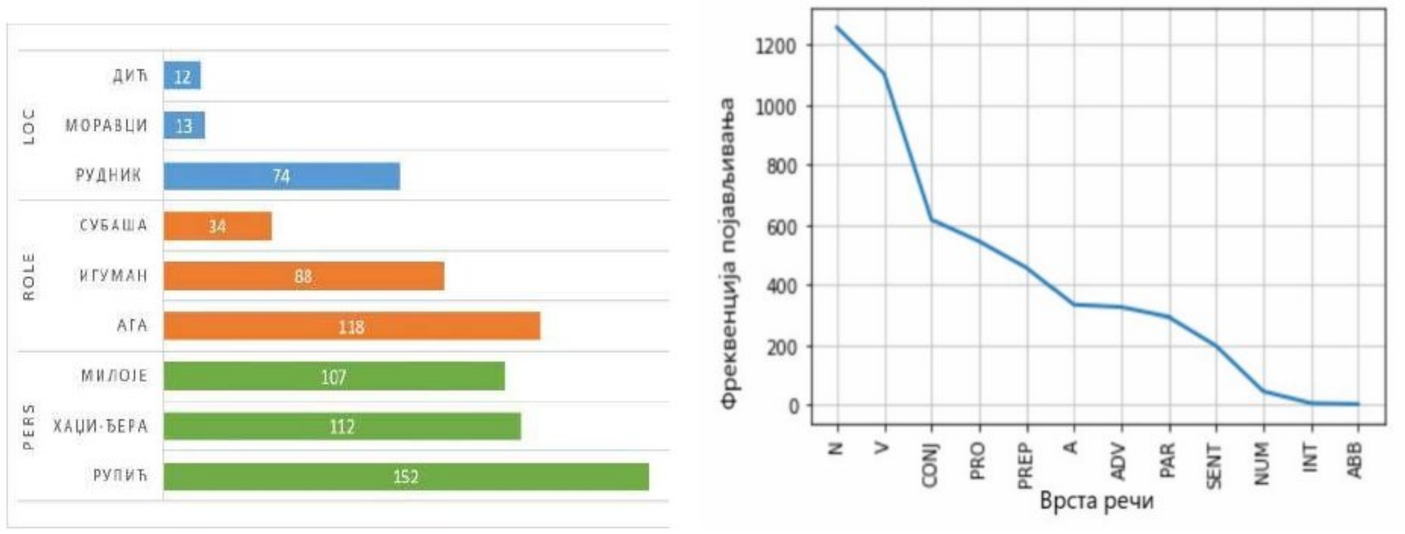
Осим припреме вишејезичног корпуса, COST акција удаљеног читања као један од циљева има анализу великих колекција књижевних текстова коришћењем рачунарских метода (Stanković et al. 2019). Анотирани корпус старих српских романа представља непроцењиву базу за тренирање модела и развијање система за препознавање именованих ентитета. Осим анотирања, интеграција обележавања врстом речи и лематизација свих речи у тексту, као и препознавање и обележавање именованих ентитета је текућа активност (Stanković et al. 2020). Као резултат, до сада је за 44 романа из SrpELTeC колекције завршено обележавање свих речи текста врстом речи, лемом и потенцијално именованим ентитетом у складу са препорукама акције. Величина до сада припремљена 44 романа за ниво 2 је мерена бројем токена[17] 2.106.410, бројем речи 1.711.250, бројем пасуса 2.740, и бројем реченица 6.647. Дистрибуција класа именованих ентитета и врста речи приказана је на слици 3. До краја акције на исти овај начин биће обрађено свих 100 романа из српске подколекције.

[17] Токени у односу на речи садрже и интерпункцијске и све друге специјалне знаке.



Слика 3. Лево: дистрибуција именованих ентитета у 44 романа SrpELTeC-а; десно: дистрибуција врста речи у истом скупу (бројеви у стубићима представљају хиљаде).

На слици 4. је приказана анализа дистрибуције именованих ентитета и врста речи у роману „Хаџи Ђера“ Драгутина Илића,[18] коришћењем алата за обраду текста и низа библиотека и програма за симболичко и статистичко обрађивање помоћу програмског језика *Python* и пакета *Natural Language Toolkit*, познатијег под називом NLTK.



Слика 4. Лево: најчешћи помињане локације, улоге (титуле) и особе у роману„Хаџи Ђера”, десно: дистрибуција врста речи у истом роману.

Један од будућих циљева је и анализа проналажења карактеристичних речи, пре свега именица у романима ELTeC колекције, као и моделирање тема које ће као резултат донети корист за сваког читаоца књиге, јер ће му омогућити лакше проналажење романа за које је заинтересован. Такође, то ће омогућити да се стекне увид у теме које преовлађују у српским романима писаним 1840-1920. године, те да се те теме пореде са темама које преовлађују у другим европским романима из истог периода.

[18] „Хаџи Ђера“ Драгутин Илић, https://udaljenocitanje.unilib.rs/pregled/22/strana/1/

5. Која је корист за све

За овакву амбициозну акцију кључ успеха представљали су њени циљеви који су били велики подстицај за све њене учеснике. Истраживачка група са Универзитета у Београду и Друштва за језичке технологије и ресурсе равноправно је учествовала у овој COST акцији, што није остало незапажено. Њихово изузетно ангажовање је омогућило развој корпуса који ће представљати значајан ресурс за разноврсна лингвистичка, филолошка, културолошка и информатичка истраживања. Овај корпус који ће садржати материјал који није обухваћен Корпусом савременог српског језика[19] биће урађен у складу са свим савременим трендовима уз употребу најсавременијих стандарда и алата. Корпус ће бити доступан како преко заједничке дистрибуције акције, тако и кроз дигиталну библиотеку Аурора[20] коју развија Јертех и која већ сада садржи и многе друге текстове који нису део ELTeC колекције. Коришћењем платформе отвореног кода *NoSketch Engine,* која је скраћена верзија програма *Sketch Engine,* корпус је слободно доступан за претраживање и анализу коришћењем језика за претрагу корпуса CQL (енгл. *Corpus Query Language*) на инстанци *NoSke*[21] Друштва за језичке ресурсе и технологије Јертех. Допуна платформи Аурора и Носке се врши повремено, са напретком развоја корпуса.

По завршетку акције, наставиће са рад на преосталих 50 романа, потенцијално и више, српске књижевности 1840-1920, а касније ће бити укључене и друге форме (краће приповетке, путописи, биографије итд.).

ELTeC колекција је већ коришћена за нека компаративна истраживања. Једно од њих је изучавање карактеристика наслова у романима из периода 1840-1920. на више европских језика: енглески, италијански, немачки, пољски, португалски, румунски, словеначки, српски, украјински, француски. Изучаване су каратеристике наслова од формалних (дужина наслова, коришћење поднаслова, експлицирање жанра) до синтаксичких и тематских (тема, ликови, место). Резултати овог истраживања јавно су доступни (Patras et al. 2020).

---

[19] Корпус савременог српског језика, http://www.korpus.matf.bg.ac.rs/korpus/login.php

[20] Аурора, http://aurora.jerteh.rs/

[21] https://noske.jerteh.rs/#dashboard?corpname=ELTeC

Извори и литература

Деретић, Јован. *Српски Роман : 1800-1950*. Нолит, 1981.

Деретић, Јован. *Историја Српске Књижевности*. Нолит, 1983.

Живан Милисавац (ур.). *Приповедачи* (vol. 66), Матица српска; Српска књижевна задруга, 1972.

Krstev, Cvetana. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade, 2008.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić and Duško Vitas. 2014: "A System for Named Entity Recognition Based on Local Grammars". *Journal of Logic and Computation*, 24(2): 473–489.

Krstev, Cvetana, and Ranka Stanković. "Old or new, we repair, adjust and alter (texts)." *Infotheca - Journal for Digital Humanities* [Online], 19.2 (2019): 61-80.

Moretti, Franco. "Conjectures on world literature." *New left review* 1 (2000): 54.

Odebrecht, Carolin, Lou Burnard and Christof Schöch (editors) *European Literary Text Collection (ELTeC)*, version 1.1.0, April 2021, COST Action Distant Reading for European Literary History (CA16204). DOI: doi.org/ 10.5281/ zenodo.4662444.

Patras, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, Berenike J. Herrmann, Cvetana Krstev, Katja Mihurko Poniž and Dmytro Yesypenko. 2020. *Dataset for ELTEC titles [Data set]*. Zenodo. http://doi.org/10.5281/zenodo.4268669

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec and Carmen Brando. 2019."Named Entity Recognition for Distant Reading in Several European Literatures." In *DH Budapest 2019*.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić and Mihailo Škorić. 2020."Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3954–3962.

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.3.0. 31st August 2021. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (7th October 2021).

Frontini, Francesca, et al. "Named entity recognition for distant reading in ELTeC." *CLARIN Annual Conference 2020*.

Cvetana Krstev
University of Belgrade
Faculty of Philology

Ranka Stanković
University of Belgrade
Faculty of Mining and Geology

Branislava Šandrih Todorović
University of Belgrade
Faculty of Philology

Milica Ikonić Nešić
University of Belgrade
Faculty of Philology

NEW TECHNOLOGIES FOR THE REVIVAL OF OLD TEXTS

Distant reading is a paradigm that involves the use of computer methods for the analysis of large collections of literary texts. In order for distant reading methods to be applied, careful selection of texts according to agreed criteria and their preparation is required. This is exactly what the COST action "*Distant Reading for European Literary History*" deals with. One of the most important goals of this action is the preparation of a multilingual, precisely balanced corpus which, when fully completed, will contain 100 novels first published in the period 1840-1920 for several European languages, including Serbian.

Keywords: distant reading, literary corpus, Serbian language processing, Part-Of-Speech annotation, lemmatization, named entities