Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920) при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза)

Научни рад DOI: 10.18485/judig.2025.1.ch12

Срђан Шућур[1], *0009-0003-6815-2887*
Јелена Марковић[2] *0000-0003-4962-3754*

Апстракт

У овом чланку представљамо пројекат Дигитализација српског књижевног наслеђа ијекавског изговора (1840-1920) при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза), односно изградњу Корпуса српског књижевног наслијеђа ијекавског изговора, који настаје као примарни резултат поменутог пројекта. У постојећим корпусима српског језика ијекавски изговор није представљен пропорционално својој заступљености код говорника српског језика као матерњег, а самим тим мање је доступан и за корпусна истраживања. Стога поменути Корпус репрезентативне књижевности писане на српском језику ијекавског изговора у изградњи, који иначе представља проширење корпуса SrpELTeC+ (из објективних разлога доминантно екавског изговора) у оквиру COST акције Distant Reading for European Literary History (COST Action CA16204), има за циљ обогаћивање и проширивање постојећих корпуса ијекавског изговора српског језика. Одредница „прва фаза“ у називу пројекта наговјештава да су планирана и будућа проширења и надоградња корпуса.

[1] Универзитет у Источном Сарајеву, Филозофски факултет Пале, Катедра за англистику, srdjan.sucur@ffuis.edu.ba
[2] Универзитет у Источном Сарајеву, Филозофски факултет Пале, Катедра за англистику

У тексту приказујемо актуелни пресјек реализације пројекта, односно процес формирања Корпуса. Представљамо одабир текстова за дигитализацију, рјешавање проблема достизања репрезентативности Корпуса, процес дигитализације одабраних текстова уз примјере рјешења, уз најаву публиковања припремљених дигитализованих облика на сајту Центра за дигиталну хуманистику при Филозофском факултету Пале, а у организацији Друштва за језичке ресурсе и технологије из Београда JePTex.

Пројекат у настајању, односно прва фаза, обухвата приповијетке Петра Кочића, Светозара Ћоровића, Васе Кондића, Марка Поповића, Милана Трифуновића, Јоаникија Памучине, Михајла Милановића, Николе Видаковића, Радована Перовића Тунгуза, Луке Грђић-Бјелокосића, Милана Трифуновића, Ђорђа Чокорила, Милке Алексић Гргурове, Љубице Поповић, Олге Ракић и Јелице Беловићеве, потом путописе Константина Хаџиристића, Јове Бесаровића, Саве Косановића, Ристе Бесаровића, Саве Пјешчића и Марка Цара, те романе Светозара Ћоровића и Радована Перовића Тунгуза.

У предстојећем периоду планирано је проширивање Корпуса ијекавског изговора књижевном критиком, кратком причом, као и другим жанровима, у циљу формирања релативно прихватљивог обима дигитализованих ресурса ијекавског изговора српског језика из разлога лингвистичке и екстралингвистичке природе.

Кључне ријечи: SrpELTeC+, дигитализација, српско књижевно наслијеђе, ијекавски изговор

1. Увод

Статус српског језика уопште у актуелном пресјеку стања у дигитализованим ресурсима може се окарактерисати на различите начине, у опсегу од релативно повољног, уважавајући критеријум обима и улоге заједнице која се њиме служи као матерњим, до релативно неповољног, по критеријуму поређења са значајнијим свјетским језицима и њиховим дигиталним ресурсима и напретком[3]. У сличним наводима и тумачењима, као норму и нехотично најчешће

[3] Када је у питању изградња корпуса српског језика у дијахронији, вриједи истаћи два ранија слична подухвата дигитализације, у основи лексикографске природе. Први, реализован у оквиру пројекта назива Речник српског језика од 12. до 18. века, који је покренут 2013. године, за циљ има дигитализовање ћириличких рукописа (~10.000 страница) Гаврила Стефановића Венцловића (1680–1749), помоћу софтвера *Transkribus*, а у сврху прикупљања грађе за Историјски речник српског језика (Polomac, et al., 2023). Други подразумијева формирање корпуса славеносрпских текстова (1750–1848), намијењеног изради Славеносрпског речника (Милановић, 2017).

узимамо екавски изговор српског језика, што, свакако, има своје друштвено-политичке и социолингвистичке разлоге. Ијекавски изговор српског језика номинално представља изговор равноправан екавском[4], али услови у којима српска говорна заједница живи, директно и индиректно, утичу на неповољан положај ијекавског изговора у процесима дигитализације српског језика.

Говорећи о статусу српског језика и књижевности у Републици Српској Бабић (2021, 74) истиче поновно актуелизовање проблема „екавице и ијекавице у српском језику – у корист екавице, с циљем језичког унифицирања Срба“, од стране одређених политичких кругова везаних за први сазив Скупштине Републике Српске. Истовремено подсјећа на *Закон о службеној употреби језика и писма у Републици Српској* (од 29. јуна 1996. године) и њиме прописану обавезу употребе екавског изговора[5]. Да се на овој обавези истрајавало, даље наглашава Бабић, били би поништени вијекови „српске ијекавске традиције“, а „одустајањем од ијекавице“[6] и сама основица српског књижевног језика би се „врло лако прогласила туђом“[7].

Једноставна претрага упита попут „ијекавски изговор српског језика“, „корпуси српског језика ијекавског изговора“, коришћењем ћириличног и латиничног писма, прилично је индикативна за статус ијекавског изговора у дигиталним изворима. Резултати таквих претрага углавном воде ка општим прегледима статуса јужнословенских језика који су објављени у Босни и Херцеговини или Хрватској, а најчешће повезују ијекавски изговор са језичким варијантама које се не одређују као српске. Међу резултатима налазимо и коментаре проф. др Душка Витаса, који без околишања каже: „ако изгубимо ијекавицу, други ће нам одузети књижевност“[8], уз навод да се:

> […] у кругу хрватских рачунарских лингвиста разматра проблем разликовања српског и хрватског језика и граница која се аутоматски подвлачи јесте екавски и ијекавски говор, а ми

[4] Предмет Паун Јовановић против Србије (Представка број 41395/15), пред Европским судом за људска права у Стразбуру слика нешто другачију екстралингвистичку реалност.

[5] У јесен 1993. године „предност екавици у језичком стандарду“ дата је одлуком руководства у Палама, но то није подразумијевало литерарни и разговорни стил (Милосављевић, 2019: 46, 51).

[6] Идеја о предности екавице у Републици Српској одбачена је 1998. године (Милосављевић, 2019: 46).

[7] О покушају подвођења свега што је ијекавско под хрватско, те подвођењу писаца свих националних опредјељења, са цјелокупног простора БиХ, који су писали ијекавицом под „босански језик“ говори Стојановић (2010: 17, 20).

[8] https://www.dnevnik.rs/kultura/vitas-ako-izgubimo-ijekavicu-drugi-ce-nam-oduzeti-knizevnost-26-06-2017

немамо добар узорак српских ијекавских говора на основу кога би било могуће конструисати моделе за програм који одлучује да ли је неки текст на српском или хрватском језику.

1.1. Корпус српског књижевног наслеђа ијекавског изговора (1840–1920) – основни подаци

Имајући у виду значајно кашњење научне јавности у формирању узорака и корпуса ијекавског изговора српског језика, као и недавно публикованих стотину романа српског језика (Krstev, 2021) у оквиру COST акције под називом „Distant Reading for European Literary History“ (Schöch, et al. 2021), односно успјех српске академске заједнице у погледу заузимања мјеста које јој и припада у тијелу романа европских језика у периоду 1840-1920, иницирана је идеја о проширивању поменутог корпуса ијекавским изговором, у фазама, користећи друге жанрове од значаја, попут приповијетки, путописа, књижевне критике, епистоларне форме, административних и других жанрова. Поменута идеја у својој првој фази, односно првом сегменту, реализује се кроз пројекат под називом *Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920) при центру за Дигиталну хуманистику Филозофског факултета Пале (прва фаза)*, који је подржан од стране Министарства за научнотехнолошки развој и високо образовање Републике Српске.

Корпус који треба да буде резултат наведеног научноистраживачког пројекта представља проширење већ објављеног корпуса SrpELTeC+[9] (Stanković, Krstev & Vitas, 2024), започетог у оквиру завршене COST акције CA16204. Овај пројекат наставак је досадашње плодне сарадње чланова пројектног тима са Друштвом за језичке ресурсе и технологије JePTex из Београда[10].

[9] Формирању српске компоненте корпуса ELTeC посвећен је тематски број часописа *Инфотека* 21(2). У овом броју описан је настанак корпуса SrpELTeC „од празне листе, до колекције од стотину романа“ (Trtovac, Milnović & Krstev, 2021, 7–25). О метаподацима придруженим овом корпусу пише Крстев (Krstev, 2021, 26–42), док процес анотације описују Станковић, Крстев, Шандрих Тодоровић и Шкорић (Stanković, Krstev, Šandrih Todorović & Škorić, 2021, 43–59).

[10] Досадашња сарадња очитава се у формирању трокомпонентног корпуса КорСАнг, који, осим превода са српског на енглески и са енглеског на српски, чине и аргументативни састави србофоних студената англистике на матерњем српском језику (КорССАнг). За више података о корпусу видјети Радоња и Шуњур (2021). Осим тога, сарадња се огледа и у значајној помоћи у изради двају докторских дисертација, од којих је једна већ успјешно одбрањена на Филозофском факултету Пале Универзитета у Источном Сарајеву (Шуњур, 2022), а друга у фази израде.

Претходно споменути корпус SrpELTeC+ је, из разлога објективне природе, који подразумијевају разне социолингвистичке факторе, те доступну грађу, доминантно екавског изговора. Његово проширење корпусом репрезентативне књижевности писане на српском језику ијекавског изговора омогућиће синхроно и дијахроно проучавање ових двају изговора српског језика, те ће представљати, између осталог, значајан искорак у приближавању потоњег изговора широј читалачкој публици, као и подстријек за будућа проширења електронских корпуса српског језика. Свакако још је важније да један од доприноса буде већа присутност ијекавског изговора српског језика у различитим облицима на глобалној мрежи, а самим тим и његово препознавање као једног од два равноправна изговора српског језика.

2. Одабир текстова за дигитализацију

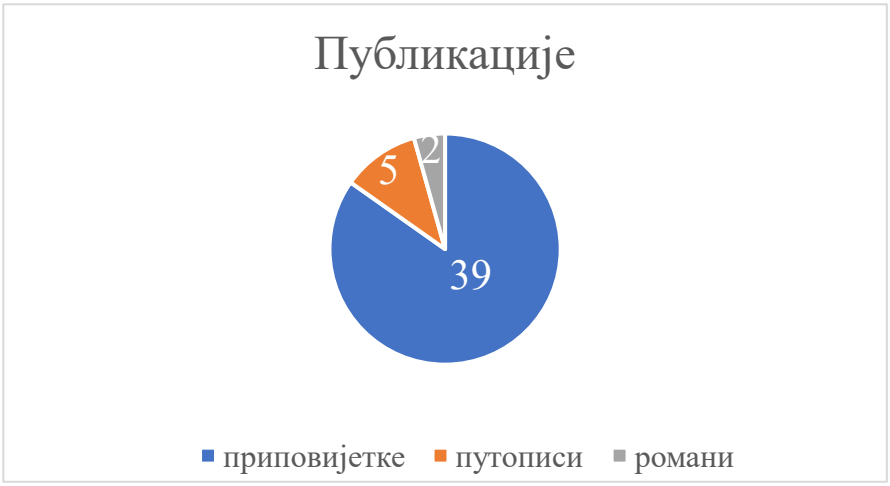
Текстуални жанрови који су планирани у првој фази јесу романи, приповијетке и, путописи. Осим романа, који су свакако циљни жанр постојећег корпуса SrpELTeC+, друга два одабрана жанра сматрали смо комплементарним жанру романа имајући у виду да су у питању жанрови писаног језика, затим да су у питању текстови који по подобности за језичка, књижевна и шира друштвено-историјска проучавања могу бити од великог значаја.

Одабир текстова за дигитализацију повјерен је члановима пројектног тима са Катедре за компаративистику, Катедре за библиотекарство и Катедре за србистику Филозофског факултета Универзитета у Источном Сарајеву. Руководећи се критеријума за израду корпуса ELTeC, у разматрање су узета како канонска, тако и мање позната или непозната дјела први пут објављена на српском језику ијекавског изговора, на ћирилици[11], у периоду од 1840–1920, те је провјерена њихова доступност у аналогном и дигиталном формату у библиотечким фондовима у земљи и окружењу.

Након што је формиран прелиминарни списак публикација које задовољавају претходно наведене критеријуме за израду корпуса, чланови тима су оцијенили значај појединачних публикација за корпус. Тиме је за дигитализацију одабрано 46 публикација, и то 1. *приповијетке*: Петра Кочића, Светозара Ћоровића, Васе Кондића, Марка Поповића, Милана Трифуновића, Јоаникија Памучине, Михајла Милановића, Николе Видаковића, Радована Перовића Тунгуза, Луке Грђић-Бјелокосића, Милана Трифуновића, Ђорђа

[11] Уз изнинмку приповијетке *Vićo*, аутора Николе Видаковића, која је објављена на латиници.

Чокорила, Милке Алексић Гргурове[12], Љубице Поповић, Олге Ракић и Јелице Беловићеве, 2. *путописи*: Константина Хаџиристића, Јове Бесаровића, Саве Косановића, Ристе Бесаровића, Саве Пјешчића и Марка Цара, и 3. *романи*: Светозара Ћоровића и Радована Перовића Тунгуза.[13] Међу публикацијама налазе се индивидуална издања, збирке приповиједака, те приповијетке објављене у периодици[14]: *Босанска вила*, *Дабро-босански источник*, *Даница*, *Зора*, *Женски свет* и *Bošnjak*. Заступљеност појединачних публикација приказана је на Графикону 1.



Графикон 1. заступљеност појединачних публикација у корпусу

3. Приповијетке

За дигитализацију су одабране сљедеће приповијетке Петра Кочића; *Гроб слатке душе*, *Мрачајски прото* и *Мргуда*, из збирке приповиједака *С планине и испод планине* (1907), те *Вуков гај*, *Змијање* и *Кроз мећаву* (*Јауци са Змијања*, 1910).

Светозар Ћоровић заступљен је приповијеткама *Мали просјак*, *Под вјештичиним кровом* и *У мајчину загрљају* (*Из моје домовине*,

[12] Ауторка је у бројевима *Босанске виле* у којима су објављена дјела одабрана за дигитализовање потписана као Милка Гргурова.

[13] Осим публикација које су ангажовањем чланова тима у пројектним активностима укључене у Корпус, придружене су и двије већ дигитализоване публикације, захваљујући сарадњи са групом JePTex из Београда.

[14] Српску периодику у Босни и Херцеговини (поред *Босанске виле* и *Зоре* ту су и *Прегледа мале библиотеке*, *Развитак* и *Српска омладина*) у вријеме аустроугарске власти истражује Ласица (2016, 2020).

1898), те приповијетком *Разорено гнијездо* (периодично објављивана у *Зори*, 1898, 1–9).

На коначни списак за дигитализовање уврштене су и приповијетке Васе Кондића; *Мајко ја идем*, *И шуме наше плачу*, *Сухо грање* (*Приповијетке*, 1910), те Марка Поповића; *Невјера*, затим *Мара, Мејрима*, и *Сеја* (*Књижевни радови Марка С. Поповића – Родољуба*[15], 1893), као и приповијетка *Све за народ*[16] (1894).

Првом фазом пројекта обухваћене су и приповијетке Милана Трифуновића; *Млади Станковићи*, *Раде*, *Чича Стојан* (*Из равне Посавине*, 1906) и *Општинска вага* (објављена у 11. броју *Женског света* 1906), затим приповијетке Михајла Милановића; *Јерко Урошевић*, *Стана*, *Христова икона* (*Босанке*, 1903), Николе Видаковића; *Vićo* (*Bošnjak*, 1904), Радована Перовића Тунгуза, објављене у *Босанској вили*; *Има љубави* (1906, 1–2), *Леле Турци* (1909, 21–22), *Тек након пет година* (1910, 1–2) и *Амајлија над амајлијама* (1911, 3–12).

По једном приповијетком заступљени су Лука Грђић-Бјелокосић; *Крвнина*, (1892) и Ђорђе Чокорило; *Крај Кошеве* (Зора, 1901).

Као и приликом формирања корпуса SrpELTec, SrpELTec-ext и SrpELTeC+, изазов је представљало проналажење женских аутора[17], те је тако у ову фазу дигитализације уврштено 8 дјела 4 различита женска аутора. За дигитализацију је одабрано пет дјела Милке Алексић Гргурове[18], која су претежно периодично објављивана у *Босанској вили*; *Двије сестре* (1897, 1–3), *Несретан случај* (1898, 14–18), *Преко љествица* (1900, 17), *Циганка* (1900, 1–3), *Три пријатеља* (1901, 13–

[15] https://digitalna.nb.rs/view/URN:NB:RS:ND_59067F874356CCC9C2256722F6BB18CC

[16] https://digitalna.nb.rs/view/URN:NB:RS:ND_D66BA3B89CF0FA0A3C2EB2F3509458CC

[17] Током отоманске власти могућности за развој образовања и културе на Централном Балкану биле су изразито ограничене. У Босни и Херцеговини у том периоду прилику да се образују имали су углавном бегови и њихове супруге. Прилике постају нешто повољније након окупације ових крајева и њихове анексије од стране Хабзбуршке монархије, која је женама понудила више основних права од оних које су уживале током Османског царства. Најсвеобухватнији приказ стања тадашње културе и образовања и њиховог значаја за жене дат је у сарајевском часопису *Српкиња* (1913)(Hawkesworth, 2000: 89, 243). Период од 1857. године (када је основана прва српска, женска, основна школа) до окупације обиљежио је просвјетни и хуманитарни рад Хаџи Стаке Скендерове (*Босанска вила*, 1903, XVIII (23/24), 393–395), а упоредо и касније и велике српске добротворке, Енглескиње Мис Аделине Паулине Ирби (енг. *Miss Adeline Paulina Irby*)(*Босанска вила*, 1904, XIX (13/14), 269).

[18] Куриозитет је да је прва приповијетка Милке Гргурове објављена у *Босанској вили* 1896. године (број 21), *Две жене*, писана екавицом, док су приповијетка *Двије сестре* (*Босанска вила*, 1897, бројеви 1-3), као и преостале приповијетке одабране за дигитализацију (у *Босанској вили* објављиване у периоду од 1897-1901), писане ијекавицом. Истовремено, њена приповијетка *У часу опасности*, објављена 1897. године у мостарској *Зори*, такође је писана екавицом.

14). По једном приповијетком у корпусу су заступљене Љубица Поповић; *Бјелка* (*Босанска вила*, 1899, 1), Олга Ракић; *Провала* (*Босанска вила*, 1899, 13–14) и Јелица Беловићева; *Зима у Кључу* (*Босанска вила*, 1907, 1).

4. Путописи

Дигитализацијом је обухваћено и пет путописа, међу којима су путопис Константина Хаџиристића *Из Босне о Босни*, уједно и најстарија публикација уврштена у прву фазу овог пројекта, периодично објављивана у *Даници* (1868, 21–34, 1869, 32–32), као и најобимнија публикација, путопис Марка Цара; *Венеција: успомене с пута* (1891). Преостала три путописа јесу *Из Стамбола у Сарајево* (*Босанска вила*, 1886, 14), Јове Бесаровића, затим *Ријеч двије са мога путовања у Цариград* (1908), Ристе Бесаровића, те путопис Саве Пјешчића *Од Требиња до Мостара*, периодично објављиван у *Добробосанском источнику* (1890, 10–12).

5. Романи

У корпус ће бити укључени и једно издање које је дигитализовано приликом формирања српске компоненте корпуса ELTeC; роман Радована Перовића Тунгуза *Из земље плача* (1906), те *Дивљак* (1913), Светозара Ћоровића[19].

6. Дигитализација корпуса

Пројектом је предвиђено одржавање пет радионица, које су реализоване у периоду од маја до октобра 2024. године, у организацији Друштва за језичке ресурсе и технологије JePTeх. Предавач на радионицама била је проф. др Ранка Станковић, а учесници чланови пројектног тима. Током радионица чланови тима су, кроз презентације, консултације и практичне задатке упознати са изазовима дигитализације и обучени за њено спровођење. Почевши од општих тема у оквиру радионице под називом „Дигитализација: од слике до дигиталног текста“, затим преко радионица званих „Обрада дигиталних текстова“, „Аутоматска анотација текстова“, „Читање, читање изблиза [корпуси, конкорданце] и читање из далека [статистичка обрада]“, уприличено је и представљање могућности база података коришћењем алата ТХМ („Текстометријска анализа текста коришћењем алата ТХМ“).

[19] https://digitalna.nb.rs/view/URN:NB:RS:ND_7C39AFC1F4EC2BD9E29F4BD666069A7F

Библиотеке које су у својим фондовима пронашле публикације одабране за дигитализацију и претходно их (или за потребе овог пројекта) скенирале, као и број појединачних публикација, наведени су у Табели 1. Три књиге преузете су са интернетског сервиса *Google Books*.

Табела 1. Библиотечки фондови из којих су преузете публикације

Фонд	Број публикација
Народна и универзитетска библиотека Републике Српске	19
Народна библиотека Србије	8
Библиотека Филозофског факултета Пале	3
Универзитетска библиотека "Светозар Марковић"	3
Народна библиотека "Радислав Никчевић" Јагодина	3
Библиотека Сарајева	3
Google Books	3
Национална и универзитетска библиотека Босне и Херцеговине	2
Библиотека Матице српске Нови Сад	1
Народна библиотека "Филип Вишњић" Бијељина	1
Укупно	**46**

Током одвијања првих трију радионица формирана је база највећег броја скенираних публикација и приступило се њиховој припреми за OCR (енг. *Optical Character Recognition*), тј. оптичко препознавање карактера, односно „рашчитавање“. Упоредо са формирањем базе ажурирани су метаподаци о публикацијама, и то: о првом издању, односно издању које ће бити дигитализовано, називу публикације, обиму, броју стубаца, жанру, аутору публикације, години рођења и смрти, периодици/збиркама приповиједака из којих су публикације/путописи издвајани, те институцијама које су помогле дигитализацију.

Припрема публикација за OCR подразумијевала је: 1) издвајање појединачних приповијетки из збирки приповиједака, 2) обједињавање приповијетки објављених периодично у јединствене документе.

Примјер приповијетке Милке Алексић Гргурове *Двије сестре* објављене у узастопна три броја *Босанске виле* (1897, 1–3) , обједињене у јединствен документ, дат је на Слици 1.

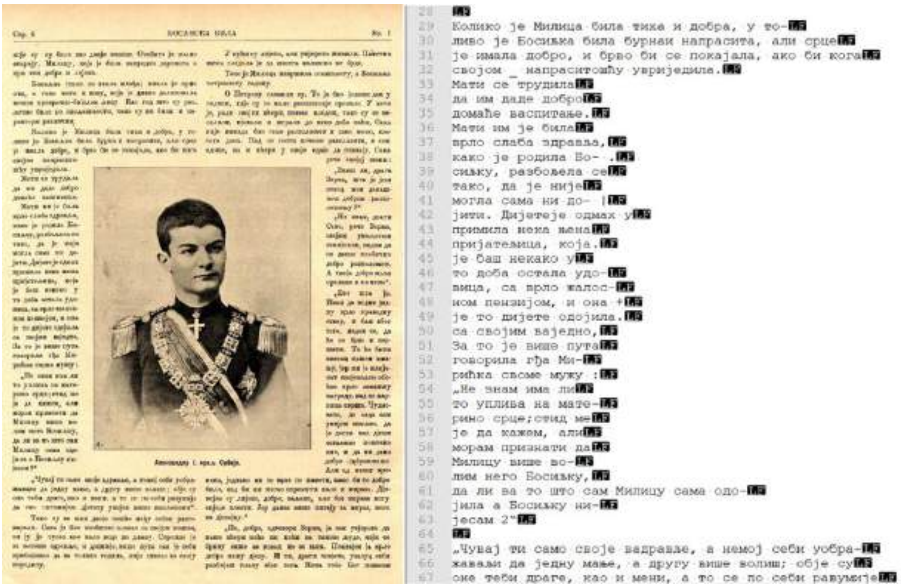
Слика 1. Приповијетка *Двије сестре*, изворно објављена у три узастопна броја *Босанске виле*, обједињена у један документ

По обављеној припреми публикација, OCR је вршен уз сарадњи са JePTex-ом, помоћу интернетског алата доступног на линку texase.jerteh.rs, заснованом на *Tesseract Open Source OCR Engine*-у[20].

Успјешност OCR-а била је условљена изворним стањем оригиналних публикација које су скениране, тј. факторима попут мрља, боје папира (посљедица старења), печата, поравнања при скенирању, биљешки накнадно писаних оловком и сл. На исход OCR-а утицала су и својства изворног текста, попут графичког уређења (нпр. централно постављене слике, в. Слика 2.) и форматирања (текст у два ступца, украси), као и накнадна обрада скенираних публикација, додавањем воденог жига (енг. *watermark*).

Након што су прикупљени OCR-овани документи, било их је неопходно додатно уредити, прије приступања аутоматској и ручној корекцији. То уређивање подразумијевало је уклањање „вишкова“; у случају публикација објављиваних периодично, било је неопходно уклонити дијелове текстова публикација које су у неком броју претходиле тексту одабраном за дигитализацију, или су га слиједиле, затим називе периодике, њихове бројеве, и сл. У приповијеткама издвојеним из збирки приповиједака уклањани су називи приповијетке/име аутора, који се у изворном тексту смјењују на парним, односно непарним страницама.

[20] https://github.com/tesseract-ocr/tesseract



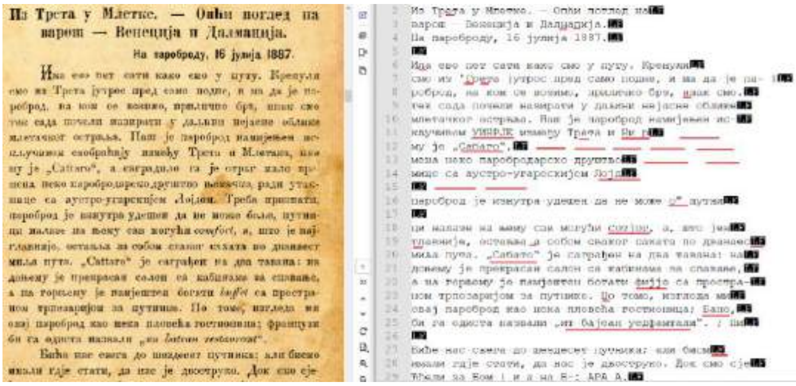
Колико је Милица била тиха и добра, у то-
ливо је Босиљка била бурнаи напрасита, али срце
је имала добро, и брзо би се покајала, ако би кога
својом _ напраситошћу увриједила.
Мати се трудила
да им даде добро
домаће васпитање.
Мати им је била
врло слаба здравља,
како је родила Во- .
сиљку, разбољела се
тако, да је није
могла сама ни до- |
јити. Дијетеје одмах у
примила нена
пријатељица, која
је баш некако у
то доба остала удо-
вица, са врло жалос-
ном пензијом, и она +
је то дијете одојила.
са својим ваједно,
За то је више пута
говорила гђа Ма-
рићка своме мужу :
„Не знам има ли
то уплива на мате-
рино срце; свиђ ме
је да кажем, али
морам признати да
Милицу више во-
лим него Босиљку,
да ли ва то што сам Милицу сама одо-
јила а Босиљку ни-
јесам ?"

„Чувај ти само своје задравље, а немој себи убра-
жавати да једну мање, а другу више волиш; обје су
оне теби драге, као и мени, а то се по себи равумије

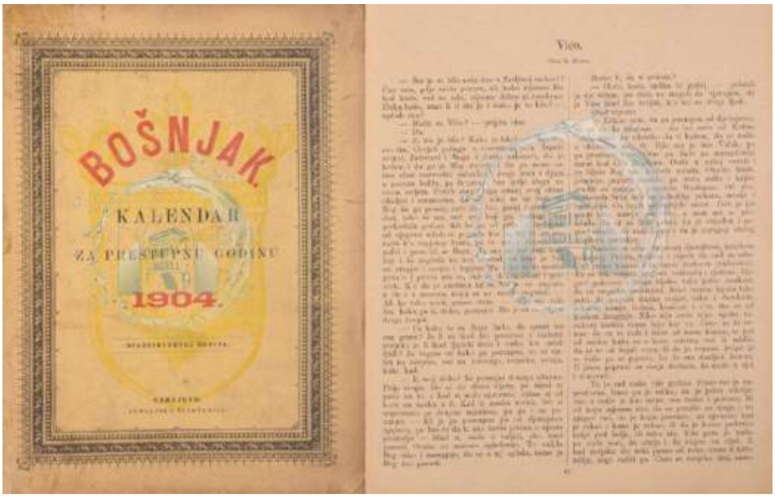
Слика 2. Примјер OCR-а текста објављеног у периодици у два ступца са централно постављеном сликом

Усљед претходно описаних фактора, у OCR-ованим документима, и поред аутоматске корекције, понекад је долазило до неразликовања ћириличних слова (м-л, т-г, и-н, м-н, п-н, в-з, с-е, ц-п), или интерпункцијских знакова (наводници, тачке, зарези, цртице и црте). У неким случајевима недостајале су ријечи (Слика 3.), или читави пасуси, а у случају једине публикације на латиници (приповијетка *Vićo*, в. Слика 4.), OCR због воденог жига није био успјешан, те је било неопходно прекуцати текст публикације. Повремено је долазило и до преклапања текста објављеног у два ступца.

Ручна корекција OCR-ованих текстова вршена је софтвером Notepad++, проширеним језицима ELTeC1 и ELTeC2, који су развијени за српски језик, помоћу електронских рјечника (Krstev, 2008), што омогућава аутоматску корекцију.



Слика 3. Примјер OCR-ованог изворног текста са недостајућим ријечима



Слика 4. Приповијетка *Vićo* и централно постављени водени жиг који је отежао OCR

Ручна корекција подразумијевала је и уређивање пасуса (тзв. „тврди крај реда“ CR LF), како би се подударали са оним у оригиналним публикацијама, те спајање ријечи које су у оригиналу прекинуте на крају реда. Паралелно са ручном корекцијом вршена је и основна анотација, тј. обиљежавање текста етикетама у складу са TEI (енг. *Text Encoding Initiative Guidelines*)(Bernard, Schöch & Odebrecht, 2021). То обиљежавање подразумијева означавање поглавља, фуснота, дијелова текста који су на страном језику, или су истакнути на неки начин (курзив, подебљана/велика слова и сл.). Коришћене етикете дате су у Табели 2.

Табела 2. TEI етикете коришћене при основној анотацији

Етикета	Означава
	почетак и крај поглавља
<head> </head>	наслов поглавља
<pb n="X[21]"/> <pb n="X.Y[22]"/>	почетак нове стране из оригинала (убацује се у непрекинут ред)
<title> </title>	назив умјетничких и других дјела
<foreign> </foreign>	текст на страном језику
<hi> </hi>	истакнути дио текста
<quote> <l> </l> </quote>	стихове пјесама
<ref target="#_NX[23]"/> <note xml:id="_NX"> </note>	фусноте

Будући да су многе публикације објављиване у периодици излазиле у неколико (узастопних) бројева, редовна пракса при приређивању часописа у периоду обухваћеном пројектом била је да се читаоцима кратким напоменама (некад испод наслова, некад на крају текста), у загради, наговијести да појединачне публикације имају наставак, или да се у конкретном броју завршавају. Неке од тих напомена (Слика 5.) биле су: (Свршиће се), (Мјесто свршетка наставак), (Продужиће се), (Наставиће се), (Наставак) и (Свршетак).


Двије сестре.
Написала Милка Грегурова.
(Мјесто свршетка наставак).
Несретан случај.
Написала Милка Грегурова.
(Наставак.)
Двије сестре.
Написала Милка Грегурова.
(Свршетак).
за прозор изнад себе.
(Наставиће се).

Слика 5. Напомене на крају текста публикација које су у периодици објављиване у више (узастопних) бројева

[21] Број стране.
[22] У случају да се ради о периодично објављиваним публикацијама, X јесте број часописа, Y број стране.
[23] У случају фуснота, X подразумијева њихов редни број.

Како би ове напомене биле сачуване и у дигитализованом облику публикација, додјељивана им је етикета <trailer> <trailer>;

<trailer>(Свршиће се).</trailer>
<trailer>(Свршетак).</trailer>,
<trailer>(Мјесто свршетка наставак).</trailer>
<trailer>(Продужиће се).</trailer> <trailer>(Наставиће се).</trailer> <trailer>(Наставак).</trailer>

Пошто путописи често обилују топонимима, ријечима и цитатима на страним језицима, те пропратним коментарима и објашњењима у виду фуснота, показали су се нарочито захтјевним током дигитализовања, што је подразумијевало уврштавање мноштва метаподатака у дигитални формат.

7. Публиковање припремљених дигитализованих облика и представљање корпуса научној јавности

У актуелном тренутку говоримо о Корпусу у настајању, имајући у виду да је завршна фаза у припреми за публиковање, у организацији Друштва за језичке ресурсе и технологије, у току. Иначе, Корпус је представљен научној јавности у току Међународне конференције *Јужнословенски језици у дигиталном окружењу (ЈуДиг)*, одржаној од 21–23. 11. 2024. године на Филолошком факултету Универзитета у Београду. Подразумијева се да ће након публиковања представљање научној јавности бити сврсисходније, будући да ће прикази укључити и конкретне примјере употребе или коришћења у начноистраживачке сврхе.

У будућим сродним подухватима планирано је проширивање Корпуса из истог периода новим жанровским компоненатама, књижевном критиком, кратким причама или информативним тектовима из доступних периодичних публикација, што би омогућило даље научноистраживачке активности базиране на ијекавском изговору, као равноправном, али запостављеном изговору српског језика. Такав истраживачки подухват имао би велики значај не само из угла лингвистичке оријентације, већ примарно екстралингвистичке, односно друштвено-историјске улоге дигитализованих публикација.

Захвалност

Овај рад настао је у оквиру пројекта под називом „Дигитализација српског књижевног наслеђа ијекавског изговора при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза)“, који суфинансира Министарство за научнотехнолошки развој и високо образовање у Влади Републике Српске (број 19.032/961-119/23, од 29. 12. 2023. године).

Овом приликом захваљујемо колегама из Народне и универзитетске библиотеке Републике Српске, Народне библиотеке Србије, Библиотеке Филозофског факултета Пале, Универзитетске библиотеке „Светозар Марковић“, Народне библиотеке „Радислав Никчевић“ Јагодина, Библиотеке Сарајева[24], Националне и универзитетске библиотеке Босне и Херцеговине, Библиотеке Матице српске у Новом Саду и Народне библиотеке „Филип Вишњић“ из Бијељине.

Прилог 1. Чланови пројектног тима који су учествовали у анотирању и кориговању текста

Члан тима	**Број текстова**	**Број страна[25]**
проф. др Јелена Марковић	4	69
проф. др Радославка Сударушић	6	115
проф. др Мирјана Лукић	8	92
доц. др Срђан Шуњур	20	292
мср Душан Пејић	3	72
мср Борјан Митровић	3	69

[24] https://digital.bgs.ba/bosanke-mihajlo_milanovic_f/

[25] Како број словних знакова на појединачним странама није униформан, обим посла који је подразумијевао дигитализовање публикација није нужно сразмјеран њиховом броју и обиму.

Извори

[1] *Босанска вила: лист за забаву, поуку и књижевност*. Сарајево. Год. XVIII, св. 23/24 (31. децембра 1903). – Год. XIX, св. 13/14 (31. јула 1904).

Литература

[1] Бабић, М. (2021). Статус српског језика и књижевности у Републици Српској. У М. Ковачевић (Ур.) *Зборник радова са Прве интеркатедарске србистичке конференције "Статус српског језика и књижевности у образовном систему"*, (63–80). Београд: Завод за унапређивање образовања и васпитања.

[2] Ласица, Б. (2016). Српски књижевни часописи у БиХ у вријеме аустроугарске власти и њихов допринос у развоју просвјете, културе и националног идентитета. У М. Летић (Прир.), *Наука и евроинтеграције*. Књига 10. Том II (55–72). Пале: Филозофски факултет.

[3] Ласица, Б. (2020). *Српска периодика у Босни и Херцеговини до 1918. године.* Пале: Српско просвјетно и културно друштво „Просвјета“.

[4] Милановић, А. (2017). Теоријско-методолошки проблеми стварања корпуса славеносрпских текстова. У Р. Драгичевић и А. Милановић (Ур.) *Научни састанак слависта у Вукове дане*, *46/3*, (137–145). Београд: Чигоја штампа.

[5] Милосављевић, П. (2019). *О обнављању србистике*. Београд: Институт за политичке студије, Матица српска у Дубровнику; Грачаница: Логос.

[6] Радоња, М. и Шућур, С. (2021). О Корпусу студената англистике (КорСАнг) и могућностима његове софтверске експлоатације. *Infotheca – Journal for Digital Humanities, 21(1)*, 37–58.

[7] Стојановић, Ј. (2010). Српски језик и државно-национални пројекти у 19. и 20. вијеку. У М. Ковачевић (Ур.). *Српски језик, књижевност, уметност - Зборник радова са међународног научног скупа одржаног на Филолошко-уметничком факултету у Крагујевцу: Језички систем и употреба језика*. Књига I, (11–30). Крагујевац: Филолошко-уметнички факултет Крагујевац, Скупштина града Крагујевца.

[8] Шућур, С. (2022). *Трансфер у фразеолошкој компетенцији у писању на енглеском језику као страном код србофоних говорника*. (Докторска дисертација). Филозофски факултет Универзитета у Источном Сарајеву.

***

[9] Bernard, L., Schöch, C. and Odebrecht, C. (2021). In Search of Comity: TEI for Distant Reading. *Journal of the Text Encoding Initiative, 14*, 1–21.

[10] Hawkesworth, C. (2000). *Voices in the Shadows. Women and Verbal art in Serbia and Bosnia.* Budapest/New York: Central European University Press.

[11] Jovanović v. Serbia, Application no. 41394/15, (2023). https://www.zastupnik.gov.rs/sr/база-праксе/пресудеодлуке-и-извршење/паун-јовановић-против-србије (датум преузимања: 7. 3. 2025.)

[12] Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Belgrade: Faculty of Philology, University of Belgrade.

[13] Krstev, C. (2021). The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata. *Infotheca – Journal for Digital Humanities*, *21(2)*, 26–42.

[14] Polomac, V., Kurešević, M., Bjelaković, I., Colić Jovanović, A., Petrović, S. (2023). Digitizing Cyrillic Manuscripts for the Historical Dictionary of the Serbian Language Using Handwritten Text Recognition Technology. *Slověne*, *12 (1)*, 295–316.

[15] Schöch, C., Patras, R., Erjavec, T. and Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, *1*, 1–19.

[16] Stanković, R., Krstev, C., Šandrih Todorović, B., Škorić, M. (2021). Annotation of the Serbian ELTeC Collection. *Infotheca – Journal for Digital Humanities*, *21(2)*, 43–59.

[17] Stanković, R., Krstev, C. and Vitas, D. (2024). SrpELTeC: A Serbian Literary Corpus for Distant Reading. *Primerjalna književnost, 47(2)*, 46–53.

[18] Trtovac, A., Milnović, V., and Krstev, C. (2021). The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection. *Infotheca – Journal for Digital Humanities*, *21(2)*, 7–25.

The Digitalisation of the Serbian Cultural Heritage of Ijekavian Pronunciation (1840–1920) with the Faculty of Philosophy Pale Digital Humanities Centre (Stage One)

Srđan Šućur, Jelena Marković

Summary

As Meša Selimović argues in "The Fortress" that what is not written does not exist, so we might say that in the contemporary digital environment what is not digitised barely exists. Thus, the conversion of the analog into the digital, i.e., digital reformatting, has certainly become the cornerstone of collecting corpora all over the world in the growing field of Digital Humanities.

Within the general wave of digital reformatting, this article is aimed at presenting the Digitalisation of the Serbian Cultural Heritage of Ijekavian Pronunciation (1840–1920) with the Faculty of Philosophy Pale Digital Humanities Centre (stage one). The basic goal of this Project is collecting and digitising the Corpus of the Serbian Cultural Heritage of Ijekavian Pronunciation (1840–1920). The Corpus is in actual fact an addition to the SrpELTeC+ corpus, collected mainly within the COST action Distant Reading for European Literary History (CA16204). The SrpELTeC+ corpus, due to objective reasons, is predominantly Ekavian, consisting of more than a hundred novels in Serbian representing the cultural heritage of the targeted period (1840-1920) within the total collection of numerous European languages and cultures.

Bearing in mind that Ijekavian pronunciation is equal to Ekavian pronunciation of the Serbian language, we state that Ijekavian is underrepresented in the currently available Serbian corpora, and is consequently unavailable for both a wider readership and potential corpus research. Having noticed the necessity to fill in the Ijekavian digital corpora vacancy, we aim to initiate the collecting of complex corpora, starting with the current Project and the Corpus under construction, of the representative literary works written in the Ijekavian pronunciation of the Serbian language of the 19th and the beginning of the 20th century.

Initially, it was proposed that the Corpus of the Serbian Cultural Heritage of Ijekavian Pronunciation (1840–1920) with the Faculty of Philosophy Digital Humanities Centre should consist of 30 publications in the first stage. Owing to subsequent contextual factors, aimed at achieving the acceptable level of representativeness, the total number of publications

has reached 46, with 39 stories, 5 travelogues and 2 novels. During the process, several challenges have been met; the uneven availability of two-decade period publications, the scarcity of women writers, the text selection criteria, and selected text digitising. The Corpus now includes both canonical and non-canonical publications, e.g., stories by Petar Kočić, Svetozar Ćorović, Vaso Kondić, Marko Popović, Milan Trifunović, Joanikije Pamučina, among others; travelogues by Konstantin Hadžiristić, Jovo Besarović, Sava Kosanović, Risto Besarović, Sava Pješčić and Marko Car; and novels by Svetozar Ćorović and Radovan Perović Tunguz.

The travelogues were the most demanding publications in the digitisation stage, owing to numerous instances of metadata to be embedded into the digitised format (place names, foreign quotations, footnotes, chapters), solved by the available TEI Guidelines. Some stories were originally published in smaller, sequential instalments in periodicals, which also required thoughtful consideration. A metadata database, containing all the relevant information about the digitised titles has been created, since its existence is one of the preconditions that make corpora usable for scientific purposes.

The Corpus is purposefully labelled "stage one", since we plan to enlarge it by future additions and expansions, using other available written genres, e.g., literary criticism and short stories. Such a collection of genre-based corpora would certainly enhance the visibility and availability of both the Serbian Cultural Heritage of the Period and Ijekavian Pronunciation of the Serbian language, which is a necessary step towards a complete chronological corpus-based description of Serbian culture, history and language.

Keywords: SrpELTeC+, digitalisation, Serbian literary heritage, Ijekavian Pronunciation