УДК 004.352.242:025.171
004.352.242:003.349
027.7(497.11):004

doi: 10.19090/cit.2020.37.35-46
Прегледни рад

Претраживе дигиталне рукописне колекције: могућност за рашчитавање српске ћирилице

Јелена Андоновски
andonovski@unilib.rs

Наташа Дакић
dakic@unilib.rs

Александра Тртовац
aleksandra@unilib.rs

Универзитетска библиотека „Светозар Марковић", Београд

Сажетак

Последњих неколико година, библиотеке и архиви, схватајући значај дигитализације, првенствено у погледу доступности садржаја, све више своју богату рукописну грађу преводе у дигитални формат како би, са једне стране, постала доступна корисницима, а са друге стране, како би је сачували од пропадања. Концепт виртуелног истраживачког окружења, настао је као део *Пројекта за препознавање и обогаћивање архивских докумената* (*Recognition and Enrichment of Archival Documents* – H2020 READ) и има потенцијал да омогући потпуно нови приступ историјским рукописним документима који се чувају у институцијама културе широм Европе. Главни циљ READ пројекта био је да се изгради виртуелно истраживачко окружење у оквиру кога би се развијале врхунске технологије за аутоматско препознавање, транскрипцију, индексирање и обогаћивање рукописних архивских докумената. Универзитетска библиотека „Светозар Марковић" се на самом почетку укључила у овај пројекат, као придружени партнер, са циљем да се развије алат који ће омогућити рашчитавање српске ћириличне рукописне грађе.

Кључне речи: библиотеке, архивска грађа, рукописна грађа, пројекат READ, Транскрибус, транскрипција, неуронске мреже, Handwritten Text Recognition – HTR, Keyword Spotting – KWS

Увод

Иновативни и узбудљиви концепт виртуелног истраживачког окружења[1] настао је као део *Пројекта за препознавање и обогаћивање архивских докумената* (*Recognition and Enrichment of Archival Documents* – H2020 READ).[2] Развојем нових револуционарних технологија он има потенцијал да омогући потпуно нови приступ историјским рукописним документима који се чувају у институцијама културе широм Европе. Последњих неколико година библиотеке и архиви, схватајући значај дигитализације првенствено у погледу доступности садржаја, све више своју богату рукописну грађу преводе у дигитални формат како би, са једне стране, постала доступна корисницима, а са друге стране, како би је сачували од пропадања.

Дигиталне библиотеке данас, поред прегледања садржаја, нуде и могућност претраге преко сваке речи у тексту, што доприноси прецизнијем и квалитетнијем одзиву у резултатима претраге. Да би се ово омогућило, већ годинама уназад се користи стандардна технологија оптичког

[1] Простор за иновативна истраживања у области развоја и коришћења технологија за аутоматско препознавање, транскрипцију, индексирање и обогаћивање рукописних докумената у сарадњи библиотекара, архивиста, истраживача из области хуманистичких наука и информатике.

[2] Recognition and Enrichment of Archival Documents, preuzeto 12. 8. 2020, http://observatory.rich2020.eu/rich/projects/view/313331.

препознавања карактера (Optical Character Recognition – OCR) у поступку дигитализације штампаног текста. Поступак OCR-а омогућава да текст буде претражив преко сваке појединачне речи, а могуће је обрадити и слике, фотографије и разне мултимедијалне садржаје. Међутим, OCR технологија развијена је за обраду штампаног текста и није ефикасна у примени на дигиталној рукописној грађи. За разлику од штампаног текста, у коме су карактери стандардизовани и нема великих одступања, рукописна грађа је по својој природи специфична. Једнако је индивидуална колико и њени аутори, зависи и од језика, скупа знакова, скраћеница, као и стилова писања који су се користили током одређеног историјског периода. Такође, у рукописима се често појављују слова која додирују горњи или доњи ред, знакови се појављују у различитим нагибима, велика слова су писана непропорционално велико и не прате линију текста, линије нису равне већ закривљене, а понекад се протежу и изван граница листа. Сваки аутор има свој начин писања, а понекад је скоро немогуће прочитати нечији рукопис. Због свих ових карактеристика неефикасно је применити OCR технологију у рашчитавању дигитализоване рукописне грађе.

Кроз пројекат READ комбиновани су врхунска истраживања, иницијативе за дигитализацију, стипендије за истраживаче у области хуманистичких наука, као и учешће заједнице корисника у служби читања, преписивања и претраживања рукописних збирки докумената. Пре само неколико година било је незамисливо да ће рачунари моћи да „читају“ историјске рукописе и да препознају и транскрибују текст старих рукописних докумената. Узимајући у обзир ова разматрања, главни циљ READ пројекта био је да се изгради виртуелно истраживачко окружење које може да окупи архивисте, библиотекаре, информатичаре, истраживаче у области хуманистике и волонтере који кроз међусобну сарадњу имају за циљ да развију врхунске технологије за аутоматско препознавање, транскрипцију, индексирање и обогаћивање рукописних архивских докумената. На овај начин створена је шанса да коришћењем специфичних знања и експертиза:

- архиви и библиотеке пруже велике дигитализоване колекције, а применом технологије заузврат добијају обогаћене дигиталне документе;
- информатичари развијају нове методе и алгоритме засноване на огромним скуповима података са референтним изворима који долазе директно из стварних студија случаја;
- стипендисти – истраживачи у области хуманистичких наука су у стању да пруже експертизу у разумевању садржаја ових рукописа;
- волонтери и корисници добијају шансу да дају свој допринос заједници на транспарентан и демократски начин.

Транскрибус

Пројекат READ реализован је у периоду од 2016. до 2019. године под финансијским покровитељством Европске комисије. У њему је учествовало 13 партнера из Европске уније, а координатор је била Група за дигитализацију и дигиталну заштиту[3] са Универзитета у Инсбруку.[4] Захваљујући претходној успешној сарадњи са овим Универзитетом, Универзитетска библиотека „Светозар Марковић“ се од почетка укључила и у нови пројекат, као придружени партнер.[5] Иако је након завршетка пројекта званично престало његово финансирање, овај концепт наставља да живи и да се развија преко новоустановљене организације назване READ-COOPSCE.[6]

[3] DEA group (Digitisation and Digital Preservation group), preuzeto 1. 8. 2020, https://www.uibk.ac.at/germanistik/einrichtungen/dea.html.

[4] Günter Mühlberger, “H2020 Project READ (Recognition and Enrichment of Archival Documents) – 2016–2019”, preuzeto 2. 8. 2020, http://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichm ent_of_Archival_Documents_-_2016-2019.

[5] Europeana, preuzeto 2. 8. 2020, https://www.europeana.eu/en.

[6] Revolutioniying Access to Handwritten Documents European Cooperative Society, preuzeto 30. 7. 2020, https://readcoop.eu/.

Од покретања пројекта READ, 2016. године, у складу са концептом стварања виртуелног истраживачког окружења, интензивно се развијају напредне технологије за рашчитавање рукописа на основу вештачких неуронских мрежа. Спроведена су бројна истраживања везана за препознавање узорака, технике рачунарског вида, анализе слике докумената, анализе распореда елемената на страници, моделовања језика и слично. Водеће истраживачке групе из ових области које су учествовале поставиле су нове стандарде у препознавању рукописног текста и проналажењу кључних речи. Развијене нове технике и алати обједињени су преко јавно доступне инфраструктуре, платформе Транскрибус.

Транскрибус је платформа за аутоматско препознавање, транскрипцију и претраживање историјских докумената[7] програмирана коришћењем апликација JAVA и SWT[8] и састоји се од експертског алата Транскрибус, веб-интерфејса[9] и неколико услуга у технологији облака (cloud technology). Основни циљ јој је стварање виртуелног истраживачког окружења и пружање подршке корисницима који се баве транскрипцијом штампаних или рукописних докумената. Превасходно је намењена истраживачима из хуманистичких наука, архивистима и библиотекарима, волонтерима и IT стручњацима, а нуди низ алата за аутоматску обраду скенираних докумената, као што су:

- препознавање рукописног текста коришћењем технологије за рашчитавање руком писаног текста (Handwritten Text Recognition – HTR),
- анализа распореда елемената на страници (Layout Analysis),
- разумевање докумената,
- идентификација писаца,
- оптичко препознавање карактера коришћењем ABBYY Finereader Engine 11.

Координатор пројекта је Група за дигитализацију и дигиталну заштиту на Универзитету у Инсбруку, која уједно и одржава Транскрибус. На њену иницијативу, током 2019. године, оформљена је организација READ-COOP SCE, чија је основна улога даљи развој платформе. Како Група заступа принципе отворене науке, већи део софтвера је у отвореном приступу, а додатне информације се могу пронаћи на страници репозиторијума GitHub.[10]

Корисницима Транскрибуса (било институционално или појединачно) на овај начин омогућено је несметано даље коришћење платформе, а могућност да рашчитавају податке из рукописних и штампаних текстова путем HTR технологије, истовремено доприноси и њеном побољшању захваљујући принципима машинског учења. Аутоматско препознавање широког спектра историјских текстова са друге стране, има значајне импликације на доступност писаних записа светске културне баштине.

Аутоматско препознавање рукописног текста или HTR (Handwritten Text Recognition) технологија

Технологија HTR функционише потпуно другачије од технологије OCR за штампане текстове.[11] Уместо фокусирања на појединачне карактере, HTR модели за препознавање рукописног текста обрађују целе речи или пак целе линије, скенирају их у различитим правцима

[7] Louise Seaward and Maria Kallio. “Transkribus: Handwritten Text Recognition technology for historical documents”, preuzeto 4. 8. 2020, https://dh2017.adho.org/abstracts/649/649.pdf.

[8] Philip Kahle et al., “Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents”, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 4 (2017): 19, preuzeto 3. 8. 2020, https://api.semanticscholar.org/CorpusID:25099654.

[9] “Transkribus”, preuzeto 2. 8. 2020, http://transkribus.eu/.

[10] GitHub, Transkribus, preuzeto 3. 8. 2020, https://github.com/transkribus/.

[11] Gundram Leifert et al., “CITlab ARGUS for historical handwritten documents”, *ArXiv* abs/1412.3949 (2014), preuzeto 3. 8. 2020, http://arxiv.org/abs/1605.08412.

и упоређују са одговарајућим дигиталним сликама.[12] На тај начин уче исправну дигиталну транскрипцију рукописних глифова.[13] Важно је напоменути да језик или алфабет којим су написани рукописи немају значајну улогу, будући да се неуронске мреже у HTR моделима не ослањају на лингвистичка знања.

Предуслов за читаву операцију јесте постојање транскрипта одређеног дела рукописног документа који се у оквиру ове платформе припрема на два начина: први је једноставно прекуцавање текста које омогућава обучавање HTR модела за аутоматско читање историјске грађе; а други је напредна транскрипција која корисницима омогућава креирање преписа који може послужити као основа за припрему дигиталног издања одабраног рукописа.

HTR модели се заснивају на алгоритмима за машинско учење, што значи да би технологија препознавања рукописног текста требало да буде обучена прекуцавањем најмање 25 страница одабраног материјала. Ово помаже машини да разуме обрасце које праве речи и карактери. Тако спремљен материјал за обуку познат је као “ground truth“.[14] Сама обука, односно транскрипција се мора извести врло темељно – у супротном, Транскрибус можда не би „учио“ онако како би требало. Транскрибус заправо „учи“ да „прочита“ рукописне текстове одређеног аутора тако што „гледа“ што више конкретног рукописа.[15]

Након избора корпуса од тридесетак страна предодређених за транскрипцију, потребно је за сваку страницу дефинисати: текстуалне блокове, линијске блокове у оквиру сваког текстуалног блока, као и основне линије на којима лежи текст. Ове информације дају оквир за транскрипцију и омогућавају разумевање редоследа читања документа. Када чита рукописни текст, читалац се често суочава са додатним текстом уметнутим између редова. Он интуитивно интегрише те додатке у свој читалачки ток, али програму као што је Транскрибус недостаје таква интуиција – потребна му је помоћ. Да би се одржао линеаран читалачки ток, потребно је унети додатне основне линије како би се интегрисао уметнути текст. Овај процес се зове сегментација и може се извршити мануелно, или уз подршку алата који анализира распоред елемената на страници, а који је интегрисан у Транскрибус. Недавна технолошка достигнућа побољшала су прецизност овог поступка, олакшавајући машинама да препознају текст у архивским документима који имају и сложеније распореде.[16] На слици 1 приказан је изглед једне странице рукописа која је прошла кроз аутоматску сегментацију.

Тек након ових припрема може се почети са процесом прекуцавања текста. Будући да је Транскрибус дизајниран да одржи транскрипцију што је могуће тачнијом, свака откривена линија рукописног текста везана је за његов еквивалент у текстуалном едитору – губитак или стварање неповезане транскрипције су скоро немогући. Такође, омогућено је и коришћење напредних функција. На пример, функција *Означавање* дозвољава транскриптору да додаје детаље одређеним ентитетима (углавном људима, местима, датумима и скраћеницама). Коришћењем ове опције, транскриптор „храни“ програм информацијама, што га у одређеном тренутку доводи до препознавања одређених речи, на пример, скраћеница „др“ за реч „доктор“. Поред поменутог *Означавања* постоји и функција *Метаподаци*, која је веома корисна за

---

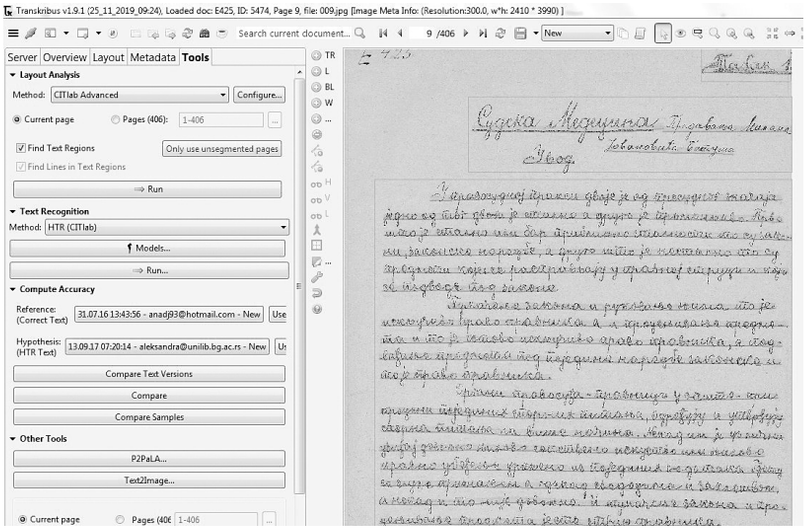
[12] Verónica Romero et al., “A Historical Document Handwriting Transcription End-to-end System”, in *Pattern Recognition and Image Analysis*, (eds) Alexandre L., Salvador Sánchez J., Rodrigues J., Vol. 10255 *Lecture Notes in Computer Science* 151 (Springer International Publishing, 2017), 149–157, DOI: https://doi.org/10.1007/978-3-319-58838-4_17.

[13] Графички облик, знак у одређеном систему писања који може бити слово, цифра, интерпункцијски или специјални знак.

[14] Basilios Gatos et al., “Ground-Truth Production in the Transcriptorium Project”, *2014 11th IAPR International Workshop on Document Analysis Systems* (2014): 239, preuzeto 2. 8. 2020, https://api.semanticscholar.org/CorpusID:12688730.

[15] Joan Andreu Sánchez et al., “Handwritten Text Recognition Competitions with the tranScriptorium Dataset”, *Document Analysis and Text Recognition* (World Scientific Publishing, 2018), 213–239, DOI: https://doi.org/10.1142/10689.

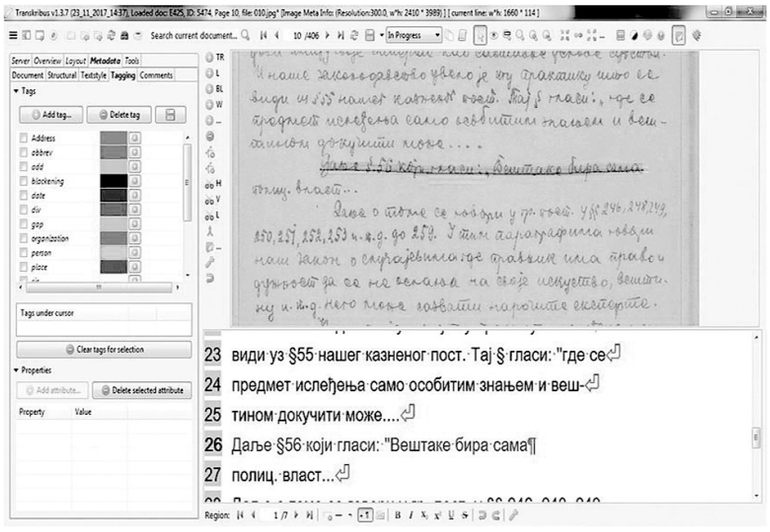
[16] Markus Diem et al., “cBAD: ICDAR2017 Competition on Baseline Detection”, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*1 (2017): 1357, preuzeto 2. 8. 2020, https://api.semanticscholar.org/CorpusID:4761833; Tobias Grüning et al., “A Robust and Binarization-free approach for text line detection in historical documents”, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*1 (2017): 238, preuzeto 3. 8. 2020, https://ieeexplore.ieee.org/abstract/document/8269978.



Слика 1. Аутоматски сегментирана страница у Транскрибусу

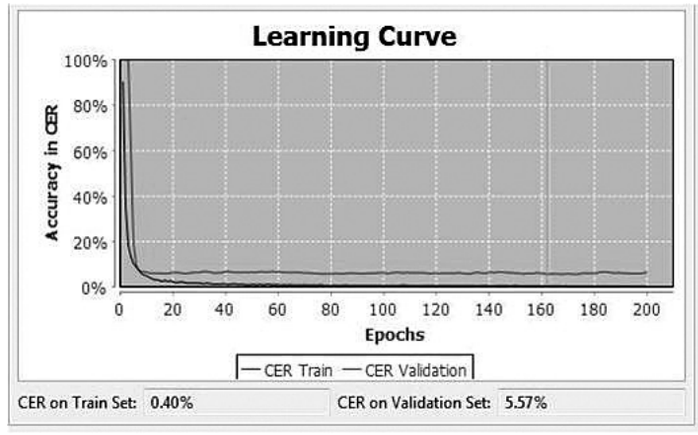
додатне стилске информације, јер омогућава транскриптору да прикаже свој препис што је могуће сличније оригиналном тексту. Ако је, на пример, аутор рукописног текста прецртао неке речи, оне се у транскрипту означавају као *Прецртано*. На тај начин, Транскрибус сазнаје да сваки ентитет који детектује као реч нужно не припада садржају писма, иако јасно припада тексту, па стога и транскрипцији. На слици 2 приказан је пример једне транскрибоване странице рукописа.

Овако припремљен (прекуцан) материјал служи за израду HTR модела, који даље омогућава аутоматско генерисање транскрипта осталих страница рукописног документа у колекцији. Основно је правило да, што је више страница припремљено и дато као материјал за обуку, то ће модел бити тачнији. Према досадашњим искуствима,потребно је најмање 15.000 преписаних речи како би се произвео добар модел. Истовремено, неопходно је из скупа припремљених страница издвојити неколико њих као скуп за тестирање са наменом процене прецизности добијеног HTR модела.



Слика 2. Транскрибована страница у Транскрибусу

Приликом израде модела, неуронске мреже упоређују знакове, речи и линије прекуцаног текста са оним што се налази на одговарајућим дигиталним сликама. Предвиђено је понављање ових задатака, а број колико пута неуронске мреже процењују дати материјал за обуку очитава се у броју задатих тзв. епоха односно итерација компарације, остварених резултата транскрипције са исправним подацима. Што више пута гледа прекуцане податке, модел се боље прилагођава специфичном рукописном стилу извора. У Транскрибусу, задата вредност за обуку HTR модела је 200 епоха.



Слика 3. Типичан пример кривуље учења HTR модела у Транскрибусу

На слици 3 приказана је кривуља учења HTR модела. На њој се може видети да, током почетних 10 или више епоха, стопа грешке карактера (Character Error Rate-CER) драстично опада и у материјалу за обуку и у скупу за тестирање. Након тога, CER значајно опада са сваком додатном епохом. Типична крива обуке неуронске мреже има, као што је и приказано, хиперболички облик. Употреба многих епоха обично доводи до снижавања CER кривуље материјала за обуку приближно на нулу, што значи да се модел потпуно прилагодио приложеном материјалу за обуку.

Недавно је у Транскрибусу развијен и имплементиран нови алгоритам заснован на напредним технологијама неуронских мрежа под називом HTR+. У поређењу са традиционалним HTR алгоритмом, HTR+ значајно скраћује време обуке модела, истовремено побољшавајући тачност препознавања текста. HTR+ модел је развио CITlab (Computational Intelligence Technology Lab) тим са Универзитета у Ростоку.[17] HTR+ користи Tensorflow[18] софтверску библиотеку коју је развио Google, чиме је омогућено да се што ефикасније конструишу дубоке неуронске мреже. Тестирања указују да је обука HTR+ модела и десетак пута бржа у односу на претходне верзије коришћених HTR алгоритма.[19] Још значајније је што се уз помоћ HTR+ технологије може аутоматски генерисати транскрипт са стопом грешке карактера (CER) до 5%. То значи да би 95% карактера у транскрипту било тачно.[20]

Ако је добијени HTR модел мање тачан, са стопом грешке карактера већом од 10%, експерименти сугеришу да тако добијена аутоматска транскрипција не би била корисна као

[17] CITlab, preuzeto 2. 8. 2020, https://www.mathematik.uni-rostock.de/forschung/projekte/CITlab/.

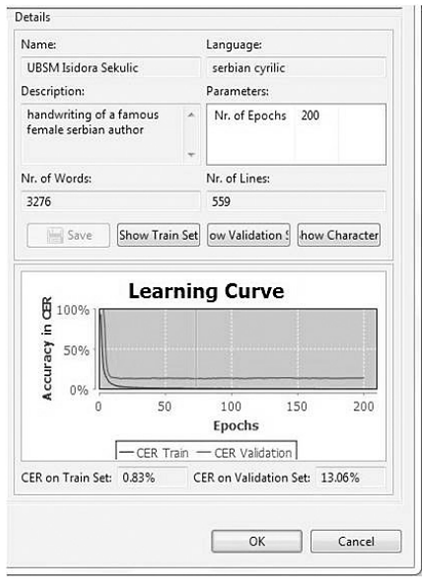
[18] Tensorflow, preuzeto 3. 8 2020, https://www.tensorflow.org/.

[19] Johannes Michael, Max Weidemann and Roger Labahn, “Deliverable 7.9, HTR engine based on neural networks P3”, Deliverable submitted to the European Commission, preuzeto 4. 8. 2020, https://read.transkribus.eu/wp-content/uploads/2018/12/Del_D7_9.pdf.

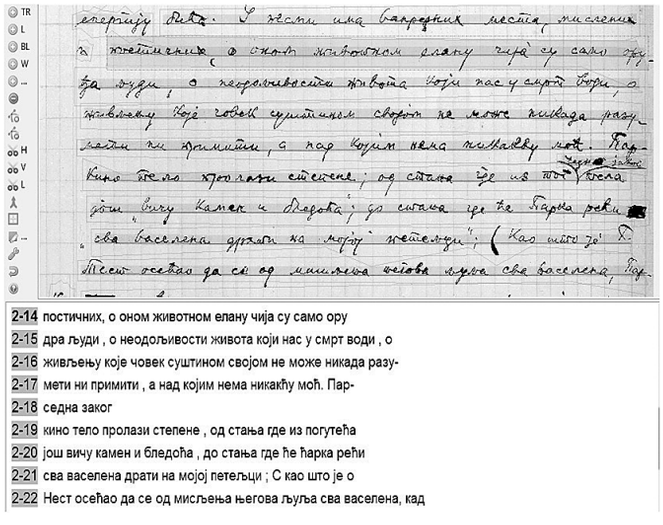
[20] Günter Mühlberger et al., “Transforming scholarship in the archives through handwritten text recognition”, *Journal of Documentation* 75 (2019): 955, preuzeto 4. 8. 2020, https://api.semanticscholar.org/CorpusID:196204627.

истраживачки ресурс, будући да исправљање безбројних грешака захтева много више времена него ручно преписивање. Међутим, из тога не следи да су мање тачни резултати на крају бескорисни. Они могу послужити као основа за прављење побољшаног HTR модела. Довољно је прекуцати још неколико страница и укључити их у претходно оформљен материјал за обуку и покренути поново израду HTR модела. На овај начин CER се може значајно смањити.

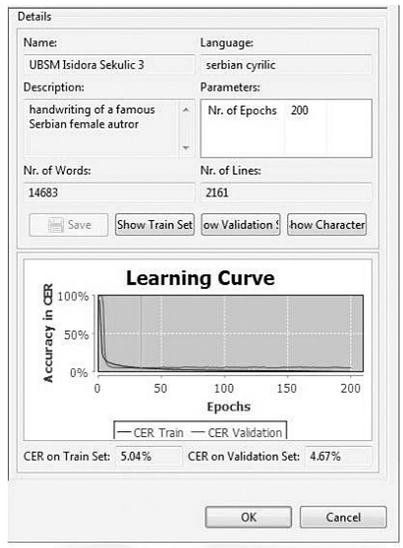
Универзитетска библиотека „Светозар Марковић“ је направила неколико HTR модела за рукописе појединих српских писаца израђених на основу рукописне грађе коју чува у својим фондовима. На приложеним примерима (слике 4, 5, 6 и 7) се могу видети добијени резултати основног и побољшаног HTR модела у програму Транскрибус за рукопис Исидоре Секулић који јасно илуструју поменута побољшања.



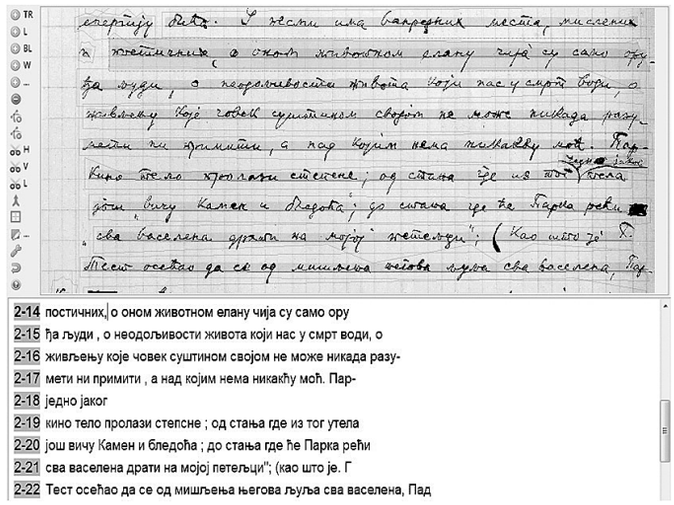
Слика 4. Процењена стопа грешке карактера за основни HTR модел припремљен на основу рукописне грађе Исидоре Секулић



Слика 5. Аутоматски генерисани транскрипт рукописа Исидоре Секулић коришћењем основног HTR модел



Слика 6. Процењена стопа грешке карактера за побољшани HTR модел припремљен на основу рукописне грађе Исидоре Секулић

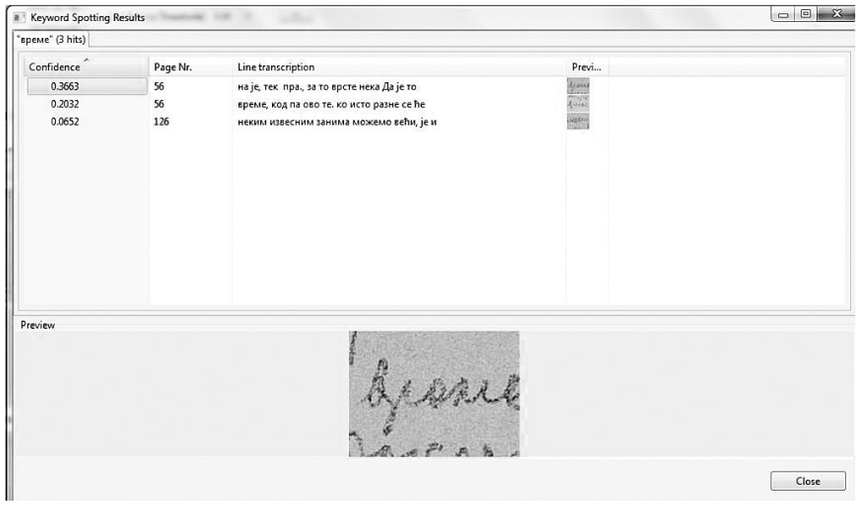


Слика 7. Аутоматски генерисати транскрипт коришћењем побољшаног HTR модела

HTR модел са стопом грешке карактера већом од 10% и даље може бити чврст темељ за претраживање и индексирање огромних колекција дигитализованих докумената. Наиме, платформа Транскрибус пружа и приступ софистицираној технологији претраживања која је позната и под називом „препознавање кључних речи“ (Keyword Spotting – KWS). Покретањем механизма претраге алат пролази кроз све вредности додељене карактерима током стварања HTR модела и враћа сва могућа подударања за задату реч. Као резулат ће бити приказано оно што програм сматра као најбоља, али ће укључити и остала могућа решења за тражену реч на основу алтернативног читања сваког знака на страници.[21] То значи да технологија

[21] Angelos P. Giotis et al., “A survey of document image word spotting techniques”, *Pattern Recognition* 68 (2017): 330, DOI: https://doi.org/10.1016/j.patcog.2017.02.023 preuzeto 2. 8. 2020, https://www.sciencedirect.com/science/article/abs/pii/S0031320317300870?via%3Dihub.

за откривање кључних речи може пронаћи речи у колекцији, чак и ако их је HTR погрешно транскрибовао. Такође, може препознати и пронаћи резултате за речи где постоје историјске или личне варијације у правопису. На слици 8 приказан је пример претраге препознавањем кључне речи „време“. Програм је као резултат дао три подударања за тражену реч. У првом резултату може се приметити да је технологија KWS пронашла реч у колекцији коју је HTR погрешно транскрибовао, уместо „време“ стоји „врсте“. Док је други резултат у потпуности прецизан.



Слика 8. Добар резултат претраге упркос погрешној транскрипцији

Рашчитавање рукописне грађе на српској ћирилици

Као што је већ речено, циљ пројекта READ био је да се направи алат који ће омогућити рашчитавање руком писаног текста како би дигиталне колекције рукописне грађе биле претраживе преко сваке појединачне речи. Од посебног интересовања за партнере на пројекту била је српска ћирилица, односно модел који ће омогућити рашчитавање рукописне грађе на српској ћирилици, што је и био основни задатак Универзитетске библиотеке „Светозар Марковић“ као придруженог партнера. У почетку је Библиотека имала задатак да се упозна са радом алата Транскрибус и транскрибује што више материјала како би се припремио “ground truth“ узорак на основу кога су колеге из Инсбрука започеле стварање HTR модела за рашчитавање српске ћирилице.

Делови програма реализовани су и кроз пројекте које је финансирало Министарство културе и информисања Републике Србије неколико година: *Нови хоризонт дигитализације* за 2016, *Рашчитана стара српска ћирилица: оживљена руком писана прошлост* за 2017. и *Рашчитаност старе српске ћирилице: историја и традиција на дохват руке* за 2018; а делимично настављени и у 2019. години када је Универзитетска библиотека, као сарадник, учествовала у реализацији пројекта *Нова парадигма архивске делатности: обезбеђивање инфраструктурних предуслова за потпуну претраживост докумената Историјског архива града Новог Сада*[22]. Кроз ове пројекте Библиотека је дигитализовала и транскрибовала значајан део рукописне грађе која се налази у њеним фондовима: рукописе Исидоре Секулић, Бранимира Ћорића, Уроша Џонића, Анице

[22] Пројекат је реализован у сарадњи са Историјским архивом града Новог Сада.

Савић Ребац, Јована Скерлића и других. За ове потребе биране су збирке рукописа које имају довољно страна руком писаног текста како би се транскрибовало што више материјала и припремио што већи “ground truth“ узорак. У оквиру пројеката припремљено је преко 4000 страна рукописног материјала, а на основу припремљеног “ground truth“ узорка креирани су модели “Cyrillic“ и “Serbian Cyrillic 20thC.“.

Од прошле, 2019. године, Библиотека је добила могућност да сама прави HTR моделе и на тај начин надограђује оно што су колеге из Инсбрука започеле, како би на крају настао кровни модел за рашчитавање српске ћирилице. Тако су направљени модели за рашчитавање рукописа поменутих аутора који би требало да постану саставни део јединственог модела. Рад је настављен и 2020. године кроз пројекат *Рашчитана стара српска ћирилица у пуном сјају: оживљено рукописно наслеђе Милоша и Михаила Обреновића*, кроз који ће се бити рашчитана рукописна грађа која се налази у фонду Универзитетске библиотеке.

Поред рукописне грађе из фонда Универзитетске библиотеке, део Транскрибуса постала је и рукописна грађа која се чува у другим библиотекама и архивима у Србији као што су: Историјски архив Крушевац, Историјски архив Суботица, Библиотека „Радоје Домановић“ у Тополи, Народна библиотека Србије, Математички институт САНУ и други. Захваљујући сарадњи са овим институцијама, дигитализована је и транскрибована рукописна грађа која представља значајан део српске културне баштине. Позване су и све друге библиотеке и архиви које имају рукописну грађу на српској ћирилици да се укључе и учествују у стварању кровног модела за рашчитавање српске ћирилице како би се створио алат који ће омогућити да дигиталне колекције рукописне грађе буду претраживе кроз цео садржај текста. У том циљу одржан је низ радионица за коришћење Транскрибуса за запослене у архивима, библиотекама и музејима широм Србије, као и велики број предавања и презентација на домаћим и регионалним скуповима библиотекара и архивиста. Акредитовани програм стручног усавршавања за библиотекаре „Демократизација дигитализације у библиотекама” посвећен је Транскрибусу и њега је већ похађало преко 388 полазника 2019. години. Програм је акредитован и за 2020. и 2021. годину.

Закључак

Технологија машинског учења омогућила је у претходних неколико година практичну примену машинског рада у областима које су до тада сматране искључиво креативним људским процесом. Примена аутоматског препознавања руком писаног текста у области очувања, проучавања и промоције културне баштине отвара нове могућности – брзу и једноставну претрагу руком писаних архивских материјала, писама и дневника, као и да се забелешке на маргинама оштећених књига учине веома брзо спремним за дигитално или штампано објављивање. Због недостатка времена и немогућности да се текст исправно прочита и прекуца, аутоматско рашчитавање мења не само опције које стоје пред истраживачима, уредницима и ауторима, већ суштински мења и приступ историјским изворима како би постали део дигиталне и реалне стварности.

Аутоматско препознавање рукописних докумената више није само замисао, већ стварна могућност. Без обзира на то да ли се ради о средњовековним кодексима или модерним архивским документима, HTR технологија не само да може да створи аутоматску транскрипцију, већ нуди и знатно побољшане опције претраживања пуног текста путем нових метода претраживања. Уз то, она омогућава и лак извоз транскрибованих докумената у различитим форматима (PDF, XML, TEI, DOCX…), чиме је омогућена њихова даља анализа.

Литература и извори:

1. CITlab. Preuzeto 2. 8. 2020. https://www.mathematik.uni-rostock.de/forschung/projekte/CITlab/.
2. DEA group (Digitisation and Digital Preservation group). Preuzeto 1. 8. 2020. https://www.uibk.ac.at/germanistik/einrichtungen/dea.html.
3. Diem, Markus, Florian Kleber, Stefan Fiel, Tobias Grüning and Basilios Gatos. “cBAD: ICDAR2017 Competition on Baseline Detection”. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*1 (2017): 1355–1360. DOI: 10.1109/ICDAR.2017.222. Preuzeto 2. 8. 2020. https://api.semanticscholar.org/CorpusID:4761833.
4. Europeana. Preuzeto 2. 8. 2020. https://www.europeana.eu/en.
5. Gatos, Basilios,Georgios Louloudis,Tim Causer, Kris Grint, Verónica Romero, Joan-Andreu Sánchez, Alejandro Héctor Toselli and Enrique Vidal. “Ground-Truth Production in the Transcriptorium Project”. *2014 11th IAPR International Workshop on Document Analysis Systems* (2014): 237–241.Preuzeto 2. 8. 2020. https://api.semanticscholar.org/CorpusID:12688730.
6. Giotis, Angelos P., Giorgos Sfikas, Basilis Gatos and Christophoros Nikou. “A survey of document image word spotting techniques”. *Pattern Recognition* 68 (2017): 310–332. DOI: https://doi.org/10.1016/j.patcog.2017.02.023. Preuzeto 2. 8. 2020. https://www.sciencedirect.com/science/article/abs/pii/S0031320317300870?via%3Dihub.
7. GitHub. Transkribus. Preuzeto 3. 8. 2020. https://github.com/transkribus/.
8. Grüning, Tobias,Gundram Leifert, Tobias Strauss and Roger Labahn. “A Robust and Binarization-free approach for text line detection in historical documents”. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 1 (2017): 236–241. Preuzeto 3. 8. 2020. https://ieeexplore.ieee.org/abstract/document/8269978.
9. Kahle, Philip, Sebastian Colutto, Günter Hackl and Günter Mühlberger. “Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents”. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 4 (2017): 19–24. Preuzeto 3. 8. 2020. https://api.semanticscholar.org/CorpusID:25099654.
10. Leifert, Gundram, Tobias Strauß, Tobias Grüning and Roger Labahn. “CITlab ARGUS for historical handwritten documents”. *ArXiv*abs/1412.3949 (2014). Preuzeto 3. 8. 2020.http://arxiv.org/abs/1605.08412.
11. Michael, Johannes, Max Weidemann and Roger Labahn. “Deliverable 7.9, HTR engine based on neural networks P3”. Deliverable submitted to the European Commission. Preuezto 4. 8. 2020. https://read.transkribus.eu/wp-content/uploads/2018/12/Del_D7_9.pdf.
12. Mühlberger, Günter. “H2020 Project READ (Recognition and Enrichment of Archival Documents) –2016–2019”. Preuzeto 2. 8. 2020. http://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019.
13. Mühlberger, Günter et al. “Transforming scholarship in the archives through handwritten text recognition”. *Journal of Documentation* 75 (2019): 954–976. Preuzeto 4. 8. 2020. https://api.semanticscholar.org/CorpusID:196204627.
14. Recognition and Enrichment of Archival Documents. Preuzeto 12. 8. 2020. http://observatory.rich2020.eu/rich/projects/view/313331.
15. Revolutioniying Access to Handwritten Documents European Cooperative Society. Preuzeto 30. 7. 2020, https://readcoop.eu/.
16. Romero, Verónica, Vicente Bosch, Celio Hernández-Tornero, Enrique Vidal and Joan-Andreu Sánchez. “A Historical Document Handwriting Transcription End-to-end System”. In *Pattern Recognition and Image Analysis*, (eds) Alexandre L., Salvador Sánchez J., Rodrigues J. Vol. 10255 *Lecture Notes in Computer Science*, 149–157. Springer International Publishing, 2017. DOI:https://doi.org/10.1007/978-3-319-58838-4_17.
17. Sánchez,Joan Andreu, Verónica Romero, Alejandro H. Toselli and Enrique Vidal. “Handwritten Text Recognition Competitions with the tranScriptorium Dataset”. In *Document Analysis and Text Recognition*, 213–239. World Scientific Publishing, 2018. DOI:https://doi.org/10.1142/10689.

18. Seaward, Louise and Maria Kallio. “Transkribus: Handwritten Text Recognition technology for historical documents”. Preuzeto 4. 8. 2020. https://dh2017.adho.org/abstracts/649/649.pdf.
19. Tensorflow. Preuzeto 3. 8 2020. https://www.tensorflow.org/.
20. “Transkribus”. Preuzeto 2. 8. 2020. https://transkribus.eu/Transkribus/.

Searchable Digitized Manuscript Collections: An Opportunity to Read Serbian Cyrillic

Summary

The READ (Recognition and Enrichment of Archival Documents) project has the potential to revolutionise access to historical collections held by cultural institutions all over Europe. This project was implemented in the period 2016/2019. It was funded by the European Commission, and involved 13 partners from the European Union. The overall objective of READ was to build a virtual research environment where archivists, humanities scholars, IT specialists and volunteers would collaborate with the ultimate goal of boosting research, innovation, development and usage of cutting edge technology for the automated recognition, transcription, indexing and enrichment of handwritten archival documents.

Since its launch in 2016, in line with its concept of creating virtual research environment, the READ project was developing advanced text recognition technology on the basis of artificial neural networks. Research in pattern recognition, computer vision, document image analysis, language modelling, but also in digital humanities, archival research and related fields has seen unprecedented progress in recent years, and European research groups are on the forefront of this specific field. Newly developed technologies and tools are integrated via publicly available infrastructure – the Transkribus platform.

The primary goal of Transkribus is to support users who transcribe printed or handwritten documents. Only a few years ago, it was still in the realm of fantasy that computers would become able to *read* historical scripts and to automatically recognise and transcribe the handwritten text of documents from the past centuries. On the other hand, users of Transkribus are able to extract data from handwritten and printed texts via HTR (Handwritten Text Recognition) technology and search digitized text without retyping, using sophisticated technology known as KWS (Keyword Spotting), while simultaneously contributing to the improvement of the same technology thanks to machine learning principles. The automated recognition of a wide variety of historical texts has significant implications for the accessibility of the written records of global cultural heritage.

Keywords: libraries, archives, manuscripts, READ project, Transkribus, transcription, neural networks, virtual research environment, Handwritten Text Recognition (HTR), Keyword Spotting (KWS)

Примљено: 24. августа 2020.
Исправке рукописа: 7. октобра 2020.
Прихваћено за објављивање: 12. октобра 2020.