821.163.41.09 Венцловић Стефановић Г.
091(=163.41):004.352.243
https://doi.org/10.18485/kij.2024.71.2.4

ВЛАДИМИР Р. ПОЛОМАЦ*
МАРКО М. МИЛОШЕВИЋ
Универзитет у Крагујевцу
Филолошко-уметнички факултет

Оригинални научни рад
Примљен: 31. 7. 2024.
Прихваћен: 24. 10. 2024.

КА (ДИГИТАЛНОМ) ИЗДАЊУ РУКОПИСА ГАВРИЛА СТЕФАНОВИЋА ВЕНЦЛОВИЋА ПОМОЋУ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ**

У раду је описан процес креирања и евалуације HTR (енгл. Handwritten Text Recognition) модела за аутоматско рашчитавање српских и српскословенских рукописа Гаврила Стефановића Венцловића помоћу софтверске платформе *Transkribus*, засноване на принципима вештачке интелигенције, машинског учења и напредних неуронских мрежа. Грађа за обуку модела узета је из два рукописа која се чувају у Архиву САНУ: *Слова изабрана*, писана српским језиком и полууставном ћирилицом; *Разглаголник*, писан српскословенским језиком и брзописном ћирилицом; као и из српскословенског *Минеја за октобар* из Архива Српске православне епархије будимске у Сентандреји, писаног полууставном ћирилицом. Најважнији резултат рада јесте креирање прве верзије генеричког HTR модела *Венцл. 1.0* са изузетним перформансама – проценат погрешно препознатих слова износи само 3%. У раду је показано како се овај модел може успешно користити и за аутоматско рашчитавање осталих Венцловићевих српских и српскословенских рукописа, што би коначно могло учинити изводљивим не само критичко издање у оригиналној графији већ и креирање дигиталних претраживих издања, а затим и електронског корпуса Венцловићеве писане заоставштине.

Кључне речи: српски језик, српскословенски језик, Гаврил Стефановић Венцловић, машинско учење, вештачка интелигенција, *Transkribus*, аутоматско рашчитавање текста, дигитално издање.

* v.polomac@filum.kg.ac.rs

** Рад је урађен у оквиру међународног билатералног пројекта *Креирање AI модела за аутоматску обраду српских средњовековних рукописа*, који под покровитељством Министарства за науку, технолошки развој и иновације Републике Србије и Немачке службе за академску размену (DAAD) заједнички реализују Катедра за српски језик ФИЛУМ-а у Крагујевцу и Институт за славистику Универзитета у Фрајбургу (Немачка).

1. Увод

Рукописна заоставштина Гаврила Стефановића Венцловића, једног од најзначајнијих српских књижевних стваралаца XVIII века, још увек нам није позната на основу целовитог критичког издања[1]. Овакво стање првенствено је условљено обимом његовог сачуваног рукописног фонда, који обухвата више од двадесет рукописа са око десет хиљада листова[2]. Основни циљ нашег прилога представља покушај да се покаже како се употребом вештачке интелигенције и машинског учења процес рашчитавања Венцловићевих рукописа може вишеструко убрзати, што би коначно могло учинити изводљивим не само критичко издање у оригиналној графији већ и креирање дигиталних претраживих издања, а затим и електронског корпуса његових српских и српскословенских рукописа.

Полазиште за настанак овога прилога представља претходно истраживање (Поломац и др. 2023) у коме је на примеру Венцловићевог рукописа *Сло-*

[1] Први избор из Венцловићевих рукописа у оригиналној графији донео је у другој половини XIX века Г. Витковић (1872, 1887). Почетком XX века Љ. Стојановић (1901) је донео каталошки опис Венцловићевих рукописа из Архива САНУ, а С. Новаковић у својим *Примерима књижевности и језика старога и српско-словенскога* (1904: 180–202) издао у оригиналној графији једно Венцловићево слово из рукописа *Великопостник*. Обиман избор из Венцловићевог стваралаштва приредио је савременом ћирилицом у другој половини XX века М. Павић (1966, 1972). Интересовање за приређивање Венцловићевих рукописа поново је оживљено од друге деценије XXI века: Т. Јовановић и Д. Стефановић (2013) приредили су фототипско издање са текстом у оригиналној графији Венцловићевог *Сентандрејског буквара* из 1717. године, М. Стефановић Бановић (2016: 270–374) донела је у оригиналној графији Венцловићеве *Беседе на Благовести* из рукописа *Слова изабрана*, а П. Пенкова (2021: 135–279) *Треће слово против аријанаца Атанасија Александријског* из Венцловићевог рукописа *Разглаголник*.

[2] Фонд Венцловићевих српскословенских рукописа нешто је обимнији од оног на српском народном језику и чине га углавном рукописи богослужбене намене: 1) *Служабник* (РГБ, Собрание Н. П. Румянцева Ф. 256 № 401), 1711–1716. године; 2) *Сентандрејски буквар* (САНУ 1 (141)), 1717. године (Јовановић/Стефановић 2013); 3) *Минеј за октобар* (СА 70 (8)), прва четвртина XVIII века; 4) *Минеј за новембар* (СА 71 (9)), прва четвртина XVIII века; 5) *Минеј за децембар* (СА 72 (10)), прва четвртина XVIII века; 6) *Минеј за јануар* (СА 73 (11)), прва четвртина XVIII века; 7) *Минеј за фебруар* (СА 74 (12)), прва четвртина XVIII века; 8) *Минеј за март и април* (СА 75 (13)), прва четвртина XVIII века; 9) *Минеј за мај* (СА 76 (14)), прва четвртина XVIII века; 10) *Минеј за август* (СА 78 (16)), 1716–1717. године; 11) *Часови и богородичник* (САНУ 78 (132)), 1725. године; 12) *Црквени зборник* (САНУ 31 (140)), 1725/30. године; 13) *Разглаголник* (САНУ 95 (135)), 1734. године; 14) *Пресађеница* (САНУ 96 (133)), 1735. године; 15) *Молитве*, *акатисти и др*. (САНУ 77 (134)), 1739. године; 16) *Каноник* (САНУ 71 (138)), 1739. године. Венцловићеву писану заоставштину на српском језику чине текстови непосредно упућени православним верницима – беседе, слова, поуке, у које понекад инкорпорира друге жанрове, попут житија: 1) *Поучења и слова разлика* (САНУ 94 (271)), 1732. године; 2) *Мач духовни* (*прва књига*) (САНУ 92 (267)), 1733/34. године; 3) *Мач духовни* (*друга књига*) (САНУ 93 (268)), 1733/34. године; 4) *Великопостник* (САНУ 97 (136)), 1740/41. године; 5) *Слова изабрана* (САНУ 101 (137)), 1743. године; 6) *Пентикости* (САНУ 98 (272)), 1743. године; 7) *Житија*, *слова и поуке* (САНУ 84 (270)), 1744/45. године; 8) *Поучење изабрано* (*прва књига*) (САНУ 99 (139)), 1745. године; 9) *Поучење изабрано* (*друга књига*) (САНУ 100 (269)), 1746. године. Поред сведочанстава о Венцловићевим рукописима који су нам остали непознати (нпр. трећа књига Барановичевих проповеди *Мач духовни* (Павић 1972: 98), овде треба поменути и претпоставку о Венцловићу као преписивачу *Богородичника* PPI28 (датиран између 1710. и 1720. године) из Библиотеке Матице српске у Новом Саду, која се износи у Грбић и др. 1999: 115–120.

ва изабрана (САНУ 101 (137))[3] показано како се помоћу софтверске платформе *Transkribus* (засноване на принципима вештачке интелигенције, машинског учења и напредних неуронских мрежа)[4] може вишеструко убрзати рад на рашчитавању грађе и креирању електронског корпуса за потребе израде историјског речника српског језика. Најважнији практични резултат рада Поломац и др. 2023 представља креирање првог HTR (енгл. Handwritten Text Recognition) модела[5] за аутоматско рашчитавање Венцловићевих рукописа на српском језику. Помоћу овога HTR модела аутоматски се могу добити прве верзије рашчитаних текстова Венцловићевих рукописа на српском језику са веома ниским процентом погрешно препознатих карактера (око 4,5–6%)[6], који варира у зависности од рукописа (детаљније у Поломац и др. 2023: 304, 307–308). Квалитативне перформансе овога модела биле су лошије од квантитативних будући да *Transkribus* није увек био у стању да препозна надредна слова која су у складу са начелима преношења грађе за историјски речник српског језика била означена као текстуални таг у суперскрипту (Поломац и др 2023: 311). Додатни проблем представљала је чињеница да је процес обуке модела био заснован на фонту *BeogradPro* који није у складу са Unicode стандардом. У ослонцу на ове резултате, у наставку истраживања желели смо да креирамо велики генерички HTR модел са унапређеним квантитативним и квалитативним перформансама способан за аутоматско рашчитавање Венцловићевих рукописа не само на српском већ и на српскословенском језику. Овај модел, заснован на фонту *Bukyvede* који је усаглашен са Unicode стандардом[7], представљао би кључни алат за вишеструко убрзавање рада на рашчитавању Венцловићевих рукописа и припреми њиховог (дигиталног) издања и електронског корпуса.

[3] Детаљније о садржају рукописа в. у Стојановић 1901: 155–171.

[4] Више о софтверској платформи *Transkribus* в. у Милбергер и др. 2019. За досадашње резултате у примени ове платформе на словенске и српске средњовековне рукописе и старе штампане књиге в. Рабус 2019, Бурлаку/Рабус 2021, Поломац/Лутовац Казновац 2021, Поломац 2022а, 2022б, 2023, Рабус/Томпсон 2023.

[5] Софтверска платформа *Transkribus* свим корисницима омогућава обуку сопственог HTR модела за аутоматско рашчитавање текста, независно од времена настанка, језика или писма. Обука HTR модела представља пример машинског учења заснованог на напредним неуронским мрежама у коме модел упоређује фотографије рукописа и одговарајућа слова, речи и линије текста у дипломатичком издању. Минимална количина података неопходна за успешно тренирање модела за старе штампане књиге износи око 5 000 речи, а за рукописе око 15 000 речи.

[6] Проценат погрешно препознатих карактера (енгл. Character Error Rate, скраћено CER) представља основни квантитативни показатељ успешности HTR модела и добија се поређењем аутоматски рашчитаног и ручно коригованог текста. HTR модел може се сматрати изузетно успешним уколико је CER испод 5%.

[7] В. https://kodeks.uni-bamberg.de/aksl/schrift/bukyvede.htm.

2. Креирање и евалуација генеричког HTR модела за Венцловићеве рукописе

Процес креирања генеричког HTR модела идеално је осмишљен за потребе аутоматског рашчитавања Венцловићевих српских и српскословенских рукописа писаних полууставном и брзописном ћирилицом. Највећи методолошки проблем за остварење овога циља представљало је одсуство јавно доступних дигиталних снимака Венцловићевих рукописа, као и ограничена финансијска средства која су нам била на располагању за потребе њихове дигитализације. Грађа за обуку модела узета је из два рукописа која се чувају у Архиву САНУ: *Слова изабрана* (САНУ 101 (137)), писана српским језиком и полууставном ћирилицом[8]; *Разглаголник* (САНУ 95 (135)), писан српскословенским језиком и брзописном ћирилицом[9]; као и из српскословенског *Минеја за октобар* (СА 70 (8)) из Архива Српске православне епархије будимске у Сентандреји[10], писаног полууставном ћирилицом.

Креирању генеричког HTR модела претходило је креирање посебних HTR модела за сва три наведена Венцловићева рукописа. У овом процесу примењен је степенасти приступ, који је посебно погодан за рашчитавање великих рукописа, а који се састоји из неколико фаза: најпре, ручно рашчитавање мањег дела рукописа (првих петнаестак хиљада речи) и обука прве верзије модела, затим, аутоматско рашчитавање следећег дела рукописа (наредних петнаестак хиљада речи), ручна корекција аутоматски рашчитаног текста и обука нове верзије модела са побољшаним перформансама. Други корак је поновљен све док квантитативне и квалитативне перформансе модела нису постале изузетне. Поступак се може илустровати на примеру рукописа *Слова изабрана* (САНУ 137):

Табела 1: Креирање HTR модела за рукопис *Слова изабрана* (САНУ 137)

HTR модел	Број страна за обуку	Број речи за обуку	CER
Венцл. САНУ 137.1	67	14 102	5,9%
Венцл. САНУ 137.2	135	29 065	5%
Венцл. САНУ 137.3	203	44 145	4%
Венцл. САНУ 137.4	270	59 091	3,51%
Венцл. САНУ 137.5	437	94 926	3%

[8] Располагали смо само дигиталним снимцима *Беседа на Божић* из овога рукописа (укупно 475 страна).

[9] Из овога рукописа били су нам на располагању само дигитални снимци *Беседа на Шестоднев* (укупно 144 страна).

[10] Располагали смо снимцима целог рукописа (укупно 440 страна) захваљујући љубазности Његовог Преосвештенства Епископа будимског, г. Лукијана.

Прва верзија модела *Венцл. САНУ 137.1* са одличним квантитативним показатељима (проценат погрешно препознатих карактера износио је 5,9%) обучена је на ручно рашчитаној грађи првих 67 страна рукописа (14 102 речи). Помоћу овога модела аутоматски је рашчитано, а затим и ручно кориговано наредних 68 страна рукописа, како би се добила додатна грађа за обуку друге верзије модела *Венцл. САНУ 137.2* са побољшаним перформансама: проценат погрешно препознатих карактера пада на 5%. Овај поступак поновљен је неколико пута, тако да је пета верзија модела *Венцл. САНУ 137.5* обучена на 437 страница текста (скоро 95 хиљада речи) са свега 3% погрешно препознатих карактера. Табела 1 потврђује претходна наша истраживања о томе да више грађе за обуку модела директно утиче на његову успешност (на мањи проценат погрешно препознатих слова) (Поломац 2022а: 16). С друге стране, на основу ове табеле, али и претходних наших искустава у обуци модела[11], може се закључити како даље повећање количине грађе за обуку не би нужно водило и даљем значајнијем унапређењу перформанси модела[12].

Степенасти приступ приликом припреме грађе за креирање модела за аутоматско рашчитавање рукописа *Слова изабрана* (САНУ 137) примењен је и на Венцловићевим српскословенским рукописима. У табели 2 приказане су перформансе модела за аутоматско рашчитавање рукописа *Разглаголник* (САНУ 135), а у табели 3 за *Минеј за октобар* (СА 8).

Табела 2: Креирање HTR модела за рукопис *Разглаголник* (САНУ 135)

HTR модел	Број страна за обуку	Број речи за обуку	CER
Венцл. САНУ 135.1	54	15 020	5 %
Венцл. САНУ 135.2	137	38 669	4%

Табела 3: Креирање HTR модела за рукопис *Минеј за октобар* (СА 8)

HTR модел	Број страна за обуку	Број речи за обуку	CER
Венцл. СА 8.1	110	23 883	4,8%
Венцл. СА 8.2	418	90 909	4%

Оба модела креирана су тако што је најпре ручно рашчитана мања количина рукописа (око 15 хиљада речи за *Разглаголник*, око 24 хиљаде речи за *Минеј за октобар*), на основу које су креиране прве верзије модела (*Венцл. САНУ 135.1* и

[11] В. модел *Dionisio 2.0* који је обучен на око 180 хиљада речи са процентом погрешно препознатих слова од 2,44% (Поломац 2022б: 159).

[12] Максималне перформансе софтвера *PyLaia* у оквиру платформе *Transkribus* у обуци модела за словенске средњовековне рукописе износе око 2–3% погрешно препознатих слова.

Венцл. СА 8.1). Помоћу ових модела аутоматски је рашчитан остатак текста, који је након ручне корекције искоришћен за обуку друге верзије са унапређеним перформансама. Поређење модела *Венцл. САНУ 135.2* и *Венцл. СА 8.2* показује како су исте перформансе (CER од 4%) постигнуте са различитом количином грађе за обуку. У случају *Минеја за октобар* за обуку модела је искоришћен цео рукопис, док је у случају *Разглаголника* могуће претпоставити да би још грађе за обуку додатно унапредило перформансе модела[13].

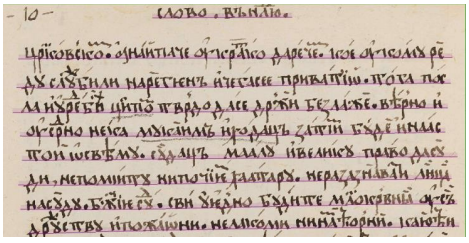
Након (полу)аутоматског рашчитавања наведена три Венцловићева рукописа могли смо да на основу њих обучимо прву верзију генеричког HTR модела под називом *Венцл. 1.0*. Параметри и перформансе овога модела приказане су у следећој табели:

Табела 4: Параметри и перформансе генеричког HTR модела *Венцл. 1.0*

Модел	Број страна за обуку	Број речи за обуку	Број епоха[14]	CER на скупу за обуку	CER на скупу за проверу
Венцл. 1.0	973	215 314	130	2%	3%

Права представа о успешности овога модела може се добити тек квалитативном анализом, односно поређењем фотографија рукописа и аутоматски рашчитаног текста. У табели 5 приказан је део листа 10а рукописа *Слова изабрана* и аутоматски рашчитан текст помоћу модела *Венцл. 1.0*.

Табела 5: Квалитативна анализа модела *Венцл. 1.0* на рукопису *Слова изабрана*

Слова изабрана (САНУ 137), део листа 10а	Аутоматски рашчитан текст, *Венцл. 1.0*
	1–1 слово. въ нл҃ю. 1–2 цр҃ковьскꙋ. а наипаче оу кра҃ко да реч҃е. ко е оу комъ ре– 1–3 дъ сл҃ꙋбили нарегиенъ и чега се е приват҃ѡ. тога пос– 1–4 ла и хр҃ѣвъ цинѡ҃ твр҃до да се др҃жи безлаже. вѣр҃но и 1–5 оу ср҃но нека мл҃каимъ и ходацъ за тїй бл҃де и надс҃– 1–6 тои ѡ свѣмъ. сл҃дацъ малъ и великъ право да сл҃ 1–7 ди, непомитъ ни почт҃е҃ хаатарь. неразазнавати лица 1–8 на сл҃дъ. бж҃іе сл҃. сви вед҃но бл҃дите мл҃окр҃вны оусъ– 1–9 дрѣствъ и пожавни. не лакоми ни нн҃ор҃ни. каюћи

Наведени узорак показује да модел *Венцл. 1.0* најчешће греши приликом препознавања размака међу речима: уместо оукра҃ко 2, слъ҃би ли 3, без лаже 4, мл҃ка

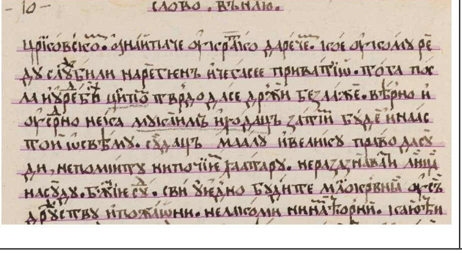
[13] Нажалост, нисмо располагали финансијским средствима за прибављање више снимака овога рукописа, тако да ову тезу нисмо успели да проверимо.

[14] Под термином *епоха* у машинском учењу подразумева „one complete presentation of the data set to be learned to a learning machine" (Бурлаку/Рабус 2021: 1).

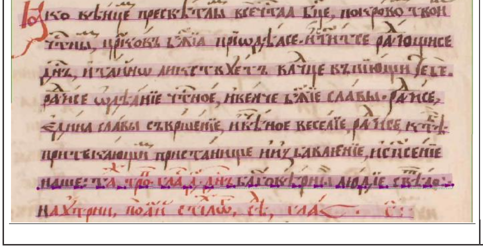
имъ 5, сꙋ– 6, не по міитꙋ 7, не разазнавати 7, ни по чїе̑ 7, оу съ– 8 модел погрешно доноси оу кра̑ко 2, слꙋ̑били 3, безлаже 4, мꙋкаимъ 5, сꙋ 6, непомiитꙋ 7, неразазнавати 7, ни почїе̑ 7, оусъ– 8. Грешке у другим категоријама долазе само изоловано: размак међу речима и надредно слово представљају проблем у примеру оу ср҆но 5 (треба оуср҆но 5); погрешно донет пајерак у примеру наас҆– 5 (треба наас– 5); погрешно донето надредно слово у примеру с̑ꙋ 8 (треба сꙋ̑ 8).

На сличне закључке упућује и квалитативна анализа аутоматског рашчитавања Венцловићевих српскословенских рукописа. У табелама 6 и 7 дат је приказ дела листа 5а рукописа *Разглаголник*, односно дела листа 10б *Минеја за октобар*, и аутоматски рашчитаног текста помоћу модела *Венцл. 1.0*.

Табела 6: Квалитативна анализа модела *Венцл. 1.0* на рукопису *Разглаголник*

Разглаголник, део листа 5а	Аутоматски рашчитан текст, *Венцл. 1.0*
	1-1 слово. въ н҃лю. 1-2 цр҃ков҆ско̑. а наипаче оу кра̑ко да рече̑. ко е оу комꙋ ре– 1-3 дꙋ слꙋ̑били нарегıенъ и чега се е приватїѡ. тога пос– 1-4 ла и ꙋре̑бѣ ципо̑ твр҆до да се др҆жи безлаже. вѣр҆но и 1-5 оу ср҆но нека мꙋкаимъ и ходацъ за тїй бꙋде и наас҆– 1-6 тои ѡ свѣмꙋ. сꙋдацъ маалꙋ и великꙋ право да сꙋ 1-7 ди, непомiитꙋ ни почїе̑ хаатарꙋ. неразазнавати лица 1-8 на сꙋдꙋ. бж҃їе с̑ꙋ. сви ꙋед҆но бꙋдите мл҃окр҆вны оусъ– 1-9 дрꙋ̑ствꙋ и пожаѡни. не лакоми ни на̑ћор҆ни. каюћи

Табела 7: Квалитативна анализа модела *Венцл. 1.0* на рукопису *Минеј за октобар*

Минеј за октобар, део листа 10б	Аутоматски рашчитан текст, *Венцл. 1.0*
	1-1 Ꙗко вѣнце̑ пресвѣтлы̑ всеч̑таа бц҃е, покрово̑ твой 1-2 ч̑тны̑, цр҃ковъ бж҃їа прїѡдѣа се. и т҃ит се ра̑чюоци се 1-3 дн҃ъ, и таинѡ̑ ликствꙋетъ вл̑чце възпїюоци тебе. 1-4 ра̑чи се ѡдѣанїе ч̑тное, и венче бж҃їе славы. ра̑чи се, 1-5 єдина славы съвршенїе, и вѣ̑ное веселче, ра̑чи се, к тѣ̑ 1-6 притекающи̑ пристаницѐ и избавленїе, и сп҃сенїе 1-7 наше: та̑, тро̑. гла̑, д҃· дн҃ъ бл҃говѣрны людїе свѣ̑ло : 1-8 на ꙋтрни, по а҃и стїлѡ̑, сѣ̑, глас· г҃ :

На основу узорка из табеле 6 може се закључити како се грешке у аутоматском рашчитавању *Разглаголника* најчешће односе на размак међу речима: уместо невѝма бѣ 1, дами 2, по водѣ 2, вьсекꙋ̑но 3 модел *Венцл. 1.0* погрешно доноси не вѝмабѣ 1, да ми 2, поводѣ 2, вьсе кꙋ̑но 3. Остале малобројне грешке односе се на писање надредног слова и титле: уместо несно̑на 6, оу̑же 8 модел погрешно доноси несnона 6, оуже 8. Грешком смо сматрали и писање пајерка у двама примерима:

конъцы 7, ѥ͡ставъникъ 8, изнад којих се на снимку види надредни знак који обликом подсећа на тачку или акценатски знак.

У узорку текста из *Минеја за октобар*, који је наведен у табели 7, најчешће грешке односе се на рашчитавање надредног слова: уместо т҃ит 2, таинѡ 3, пристаниџе 6 модел погрешно доноси: т̃ит 2, таинѡ̃ 3, пристаниџе̃ 6. Пајерак је изостављен у примеру съвръшенїе 5 (треба съвр҃шенїе 5), а титла је погрешно донета у примеру дн̃ъ 3 (треба дн̃ъ 3). У једном примеру погрешно је рашчитано слово ї: уместо веселїе 5 модел погрешно доноси веселүе 5.

3. Примена модела *Венцл. 1.0* на друге Венцловићеве рукописе

Претходно истраживање (Поломац и др. 2023: 307–310) показало је да се помоћу модела обученог на материјалу Венцловићевог рукописа *Слова изабрана* аутоматски могу добити рашчитани текстови и других Венцловићевих рукописа на српском језику са одличним квантитативним и квалитативним перформансама. У наставку истраживања интересовале су нас квантитативне перформансе генеричког HTR модела *Венцл. 1.0* на другим Венцловићевим рукописима писаним српским и српскословенским језиком. За ове потребе креиран је експеримент у коме су помоћу модела *Венцл. 1.0* аутоматски рашчитани мањи узорци (првих десет страна) српских рукописа *Великопостник* (САНУ 136) и *Поученије изабрано* (*прва књига*) (САНУ 139), као и српскословенских рукописа *Пресађеница* (САНУ 133) и *Минеј за новембар* (СА 9). Након ручне корекције аутоматски рашчитаних текстова, израчунат је проценат погрешно препознатих карактера (CER) за све рукописе, приказан у следећој табели:

Табела 8: Примена модела *Венцл. 1.0* на друге рукописе

Рукопис	CER
Великопостник (САНУ 136)	3,13%
Поученије изабрано (САНУ 139)	2,66%
Пресађеница (САНУ 133)	4,29%
Минеј за новембар (СА 9)	3,80%

Наведена табела показује како се HTR модел *Венцл. 1.0* може ефикасно користити и за аутоматско рашчитавање других Венцловићевих српских и српскословенских рукописа. Изузетне перформансе модела на рукопису *Поученије изабрано* (САНУ 139) чак су и незнатно боље него на рукописима на којима је модел обучен, док су на осталим рукописима сличне или тек незнатно слабије од процента погрешно препознатих карактера на скупу података за проверу мо-

дела. Овакви квантитативни показатељи очекивани су с обзиром на чињеницу да су наведени рукописи по језику и писму веома слични рукописима на којима је модел обучен[15]. Квалитативна анализа перформанси модела *Венцл. 1.0* на рукописима из претходне табеле није указала на битнија одступања у односу на стање наведено уз табеле 6 и 7. И на овим рукописима модел најчешће греши приликом препознавања размака међу речима, ређе приликом препознавања надредних слова, титле и пајерка.

4. Закључне напомене

Истраживање проведено у раду показало је како се помоћу софтверске платформе *Transkribus*, засноване на принципима вештачке интелигенције, машинског учења и напредних неуронских мрежа, може креирати генерички HTR (енгл. Handwritten Text Recognition) модел са изузетним показатељима (проценат погрешно препознатих карактера износи свега 3%) у аутоматском рашчитавању обимне писане заоставштине Гаврила Стефановића Венцловића. Овај генерички HTR модел, назван *Венцл. 1.0*, а обучен на грађи од 973 стране и око 215 хиљада речи коју чине део рукописа *Слова изабрана* (српски језик, полуустав), део рукописа *Разглаголник* (српскословенски језик, брзопис) и цео рукопис *Минеј за октобар* (српскословенски језик, полуустав), може се ефикасно употребити и за аутоматско рашчитавање других Венцловићевих српских и српскословенских рукописа. Квалитативна анализа перформанси овога HTR модела показала је да се најчешће грешке односе на препознавање размака међу речима, ређе на препознавање надредних слова, титле и пајерка, а само изузетно на препознавање појединачних слова у реду. Помоћу HTR модела *Венцл. 1.0* процес рашчитавања Венцловићевих рукописа може се десетоструко убрзати, чиме се коначно стварају претпоставке за припрему целовитог критичког издања у оригиналној графији његове писане заоставштине не само у традиционалном штампаном већ и у дигиталном формату. Предуслов за реализацију овога прворазредног националног и културног задатка јесте завршетак дигитализације Венцловићевих српских и српскословенских рукописа у Архиву САНУ.

[15] Интересантно би било истражити перформансе овога модела на Венцловићевим српским рукописима писаним брзописном ћирилицом (нпр. *Мач духовни*) будући да они нису били обухваћени обуком модела. Претпоставка о нешто лошијим перформансама у односу на рукописе из табеле 8 није могла бити проверена услед одсуства дигиталних снимака рукописа.

ЛИТЕРАТУРА

Бурлаку/Рабус 2021: C. Burlacu, A.Rabus, Digitising (Romanian) Cyrillic using Transkribus: new perspectives, *Diacronia*, 14, 1–9.

Витковић 1872: Г. Витковић, О књижевном раду јеромонаха Гаврила Стефановића, *Гласник Српског ученог друштва*, 24, 151–177.

Витковић 1887: Г. Витковић, *Прошлост, установа и споменици угарских краљевих шајкаша: од 1000 до 1872*, У Београду: У Штампарији Краљевине Србије.

Грбић и др. 1999: Д. Грбић, К. Шкорић, М. Гроздановић-Пајић, *Ћирилске рукописне књиге Библиотеке Матице српске. Књ. 7: Акатисти, Стихологије, Богородичници*, Нови Сад: Библиотека Матице српске.

Јовановић/Стефановић 2013: Т. Јовановић, Д. Стефановић, *Венцловићев Сентандрејски буквар: 1717*, Будимпешта–Београд: Радионица „Венцловић"–Арте.

Милбергер и др. 2019: G. Mühlberger, L. Seaward, M. Terras, S. Oliveira Ares, V. Bosch, M. Bryan, S. Colluto, H. Déjean, M. Diem, S. Fiel, B. Gatos, A. Greinoecker, T. Grüning, G. Hackl, V. Haukkovaara, G. Heyer, L. Hirvonen, T. Hodel, M. Jokinen, P. Kahle, M. Kallio, F. Kaplan, F. Kleber, R. Labahn, M. Lang, S. Laube, G. Leifert, G. Louloudis, R. McNicholl, J. Meunier, J. Michael, E. Mühlbauer, N. Philipp, J. Pratikakis, J. Puigcerver Pérez, H. Putz, G. Retsinas, V. Romero, R. Sablatnig, J. Sánchez, P. Schofield, G. Sfikas, C. Sieber, N. Stamatopoulos, T. Strauss, T. Terbul, A. Toselli, B. Ulreich, M. Villegas, E. Vidal, J. Walcher, M. Wiedermann, H. Wurster, K. Zagoris, Transforming scholarship in the archives through handwritten text recognition, *Journal of Documentation*, 5/75, 954–976.

Новаковић 1904: С. Новаковић, *Примери књижевности и језика старога и српско-словенскога*, Београд: Краљевска-српска државна штампарија.

Павић 1966: М. Павић, *Црни биво у срцу*, Београд: Просвета.

Павић 1972: М. Павић, *Гаврил Стефановић Венцловић*, Београд: Српска књижевна задруга.

Пенкова 2021: П. Пенкова, *Буквалният превод на Трето слово против арианите от Атанасий Александрийски по ръкописа на Гаврило Стефанович Венцлович в зборника Разглаголник 1732–1734. г.: Изследване и издание на текста*, София: Издателство „Валентин Траянов".

Поломац/Лутовац Казновац 2021: V. Polomac, T. Lutovac Kaznovac, Automatic Recognition of Serbian Medieval Manuscripts by Applying the Transkribus Software Platform: Current State and Future Perspectives, *Зборник Матице српске за филологију и лингвистику*, LXIV/2, 7–26.

Поломац 2022а: V. Polomac, Serbian Early Printed Books from Venice: Creating Models for Automatic Text Recognition using Transkribus, *Scripta&e-Scripta*, 22, 11–29.

Поломац 2022б: V. Polomac, Serbian Early Printed Books: Towards Generic Model for Automatic Text Recognition using Transkribus, in: D. Fišer, T. Erjavec (eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana: Institute for Contemporary History, 154–161.

Поломац 2023: V. Polomac, Macarius: A HTR model for Romanian Slavonic Early Printed Books, *Slavistica Vilnensis*, 68/2, 10–23.

Поломац и др. 2023: V. Polomac, M. Kurešević, I. Bjelaković, A. Colić Jovanović, S. Petrović, Digitizing Cyrillic Manuscripts for the Historical Dictionary of the Serbian language using Handwritten Text Recognition Technology, *Slověne*, 12/1, 295–316.

Рабус 2019: A. Rabus, Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach using Transkribus, *Scripta&e-Scripta*, 19, 9–32.

Рабус/Томпсон 2023: A. Rabus, W. Thompson, Performance of Generic HTR Models on Historical Cyrillic and Glagolitic: Comparison of Engines, *Scripta&e-Scripta*, 23, 11–34.

Стефановић Бановић 2015: М. Стефановић Бановић, *Беседе, слова и поуке на Благовести у преводу и преради Гаврила Стефановића Венцловића. Докторска дисертација у рукопису*, Београд: Филолошки факултет.

Стојановић 1901: Љ. Стојановић, *Каталог рукописа и старих штампаних књига Српске краљевске академије*, Београд: Српска краљевска академија.

Vladimir R. Polomac
Marko M. Milošević

TOWARDS A (DIGITAL) EDITION OF GAVRILO STEFANOVIĆ VENCLOVIĆ'S MANUSCRIPTS USING ARTIFICIAL INTELLIGENCE

Summary

The paper describes the process of creating and evaluating an HTR (Handwritten Text Recognition) model for the automatic text recognition of Serbian and Serbian Church Slavonic manuscripts by Gavrilo Stefanović Venclović using the Transkribus software platform, based on the principles of artificial intelligence, machine learning, and advanced neural networks. The training material for the model was taken from two manuscripts held in the Archives of the Serbian Academy of Sciences and Arts: *Slova izabrana*, written in the Serbian language and semi-uncial Cyrillic script; *Razlagolnik*, written in the Serbian Church Slavonic language and cursive Cyrillic script; and the Serbian Church Slavonic *Menaion for October* from the Archives of the Serbian Orthodox Diocese of Buda in Szentendre, written in semi-uncial Cyrillic script. The most important result of the work is the creation of the first version of the

generic HTR model *Vencl. 1.0* with exceptional performance – the percentage of incorrectly recognized characters is only 3%. The paper demonstrates how this model can be successfully used for the automatic transcription of other Serbian and Serbian Church Slavonic manuscripts by Venclović, which could finally make feasible not only a critical edition in the original script but also the creation of searchable digital editions, and subsequently, an electronic corpus of Venclović's written legacy.

Keywords: Serbian language, Serbian Church Slavonic, Gavrilo Stefanović Venclović, machine learning, artificial intelligence, Transkribus, automatic text recognition, digital edition.