Umelá inteligencia pomáha sprístupňovať písomné dedičstvo

Artificial intelligence helps to access manuscript heritage

Prof. PhDr. Dušan Katuščák, PhD. / Ústav bohemistiky a knihovnictví, Filozoficko-přírodovědecká fakulta v Opavě, Slezská univerzita (Department of Czech Studies and Librarianship, Faculty of Arts and Sciences in Opava, Silesian University), Masarykova třída 343/34, 746 01 Opava. Štátna vedecká knižnica v Banskej Bystrici (State Scientific Library in Banská Bystrica), Lazovná 240/9, 975 58 Banská Bystrica, Slovensko [1,2]

Resumé:
Témou prípadovej štúdie je vedecký a metodologický kontext európskeho projektu základného výskumu READ a aplikácia výsledkov tohto výskumu na Slovensku a v Česku. Autor upozorňuje na pokračovanie projektu READ a pokrok vo výskumoch, aplikáciách a experimentoch, ktorým sa venuje medzinárodná komunita *digital humanities* v rámci združenia READ-COOP od roku 2019. Súčasťou týchto aktivít je aj slovenský projekt aplikovaného výskumu a grantu s akronymom SKRIPTOR, rozplánovaný na roky 2020–2024. Na základe informačného prieskumu a výberu najnovšej literatúry ukazuje pokrok vo výskume a aplikáciách v oblasti optického rozlišovania písma OCR. Jadro štúdie je zamerané na používateľský a nie informatický prístup k využitiu platformy *Transkribus* na automatické rozpoznávanie textov historických dokumentov. Popisuje skúsenosti a poznatky získané pri osvojovaní si platformy *Transkribus*, ktorá využíva umelú inteligenciu stroja OCR a metódu HTR+. V štúdii sú vysvetlené a ilustrované jednotlivé hlavné kroky experimentov, proces učenia stroja až po vytvorenie nových modelov transkripcie a výsledkov automatickej transkripcie tlačenej fraktúry a rukopisných listov Andreja Kmeťa. Štúdia predstavuje aj prvý nový efektívny model transkripcie historického tlačeného písma slovenskej fraktúry (švabachu). Najprv vysvetľuje unikátny experiment s transkripciou tlačených slovenských a českých textov fraktúry. Nasleduje popis pokročilej experimentálnej transkripcie rukopisných listov Andreja Kmeťa. Predstavuje možnosti sprístupnenia transkribovaných zbierok a dokumentov v lokálnych sieťach a na internete.

Klíčová slova: *digital humanities, OCR, READ-COOP, umelá inteligencia, platforma Transkribus, HTR+, projekt SKRIPTOR, Andrej Kmeť, švabach, fraktúra, antikva, read & search*

Summary:
The topic of the study is the scientific and methodological context of the European project of basic research READ and application of the results of this research in Slovakia and the Czech Republic. The study is part of the ongoing applications of the READ project. It shows the progress of research, applications and experiments undertaken by the digital humanities international community involved in the READ-COOP association since 2019. Part of these activities is also a Slovak project of applied research with the acronym of SKRIPTOR, planned for 2020-2024. Based on information survey and selection of the latest information sources, there has been some progress in research and applications in the field of OCR. The core of the study is focused on the user-centred rather than IT-based approach to the use of the Transkribus platform for automatic text recognition of historical documents. It describes the experience and knowledge gained in adopting the Transkribus platform that uses artificial intelligence of the OCR machine and the HTR+ method. The study explains and illustrates the main steps of the experiments, the process of training of the machine, the creation

[1] ORCID: 0000-0001-7444-1077. Slezská univerzita Opava. Filozoficko-přírodovědecká fakulta v Opavě; Ústav bohemistiky a knihovnictví. Štátna vedecká knižnica v Banskej Bystrici.

[2] Štúdia je výstupom z riešenia projektu APVV-19-0456 SKRIPTOR – *Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov.*

of new models of transcription, and the results of automatic transcription of printed Fraktura texts and manuscripts by Andrej Kmeť. The study also presents the first new efficient transcription model for printed historical type of Slovak Fraktur (Gothic) script in the Transkribus platform. First, it explains a unique experiment with the transcription of printed Slovak and Czech Fraktur texts. This is followed by a description of the advanced experimental transcription of Andrej Kmeť's handwritten letters. It presents the possibilities of making transcribed collections and documents available on local networks and on the Internet.

Keywords: *digital humanities, OCR, READ-COOP, artificial intelligence, Transkribus platform, HTR+, SKRIPTOR project, Andrej Kmeť, schwabacher, fraktur, antiqua, read & search*

Úvod

Najvýznamnejší pokrok vo výskume, vývoji a aplikáciách v digitalizácii v spoločenských a humanitných odboroch, čiže v *digital humanities*, nastal najmä v posledných desiatich rokoch. Predmetom odborného záujmu je automatické optické rozlišovanie písma (OCR)[3]. Kým OCR bežných tlačených dokumentov je už dávnejšie dostatočne zvládnuté pomocou kvalitných nástrojov OCR, tak náročnejšej problematike transkripcie historických rukopisov a tlačí s využitím umelej inteligencie sa venujú desiatky výskumníkov a experimentátorov len v posledných rokoch. Pokrok nastal realizáciou projektu READ[4], ktorý ako vedecký projekt základného výskumu podliehal priamo Európskej komisii a bol ročne hodnotený nezávislými hodnotiteľmi[5]. Rozvíjajú sa aj iné platformy, aplikácie a nástroje transkripcie. Hlavným výstupom projektu READ je použiteľná platforma a nástroj *Transkribus*[6], ktoré predstavujú svetovú inováciu zameranú na transkripciu historických rukopisov a dokumentov. V strednej a východnej Európe je Slovensko zatiaľ jedinou krajinou, ktorá sa usiluje rozpracovať podnety Európskeho základného výskumu READ v projekte aplikovaného výskumu SKRIPTOR[7].

---

[3] OCR – Optical Character Recognition (Optické rozlišovanie písma)

[4] READ Recognition and Enrichment of Archival Documents, projekt, ktorého riešenie prebiehalo v rokoch 2016–2019 v rámci programu Horizon2020. [cit 2. 10. 2021]. Dostupné z: *https://cordis.europa.eu/project/id/674943*.

[5] Dušan Katuščák bol jedným z troch hodnotiteľov projektu READ pre Európsku komisiu.

[6] *Transkribus.* Komplexná platforma na digitalizáciu, rozpoznávanie textu podporované umelou inteligenciou, ako aj na prepis a vyhľadávanie historických dokumentov – z akéhokoľvek miesta, kedykoľvek a v akomkoľvek jazyku. V *Transkribus Lite* je možné použiť zbierky *Transkribus Expert Client* v prehliadači osobných počítačov a smartfónov. Mnohé z funkcií od klienta *Transkribus Expert Client* môžu byť použité aj v *Transkribus Lite*. Platforma integruje nástroje vyvinuté výskumnými skupinami v celej Európe, vrátane skupiny Pre rozpoznávanie vzorov a technológie ľudského jazyka Technickej univerzity vo Valencii a skupiny CITlab University Rostock. V októbri 2022 mal Transkribus viac ako *94 000 používateľov*, *40 mil. obrazov*, *20 mil. rozpoznaných strán*. Platforma bola vytvorená v kontexte dvoch projektov EÚ *tranScriptorium* (2013–2015) a READ (2016–2019).

[7] SKRIPTOR. Projekt APVV-19-NEWPROJECT-17816 (2020–2024). Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov [*Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts*]. Riešiteľské organizácie: Univerzita Mateja Bela v Banskej Bystrici (zodpovedný riešiteľ doc. Imrich Nagy, PhD.); Štátna vedecká knižnica v Banskej Bystrici – partner (garant prof. PhDr. Dušan Katuščák, PhD.).

Digital humanities a projekt READ

Digital humanities považujeme za spoločné pomenovanie a prierezovú metodológiu pre všetky aplikácie informačných a komunikačných technológií v spoločenských a humanitných vedách, odboroch a disciplínach a im zodpovedajúcej praxi. Táto metodológia sa komplexne uplatnila v projekte READ, ktorý sa realizoval v rámci programu Horizon 2020[8]. Autorom a koordinátorom projektu bol prof. G. Mühlberger z Univerzity v Insbrucku. Projekt READ bol financovaný Európskou úniou sumou približne 8,2 milióna EUR. Financovanie sa skončilo 30. 6. 2019. Univerzita v Innsbrucku od roku 2016 skúma základné technológie *segmentácie* textu, *rozpoznávania* rukopisu, vyhľadávanie *kľúčových slov* pre historické dokumenty a nástroje *sprístupnenia* výsledkov. Na všetkých oblastiach výskumu sa podieľali tímy univerzít vo Valencii, Rostocku, Technickej univerzity vo Viedni a ďalšie výskumné inštitúcie. Rozvinula sa spolupráca s ďalšími partnermi z 27 krajín. Výskum a vývoj naďalej prebieha. Tisíce používateľov platformy *Transkribus* tvoria nové modely transkripcie na základe historických rukopisných a tlačených zbierok národných inštitúcií, najmä knižníc a archívov. Spolupráca s komunitou výskumníkov sústredených okolo platformy *Transkribus* môže byť užitočná pre české a slovenské prostredie odborníkov z *digital humanities*.

Spoločnou víziou vedcov, expertov a iných používateľov je, aby sa verejne dostupné modely transkripcie postupne stali užitočným spoločným nástrojom pre automatickú transkripciu historických dokumentov. Je potrebné dosiahnuť takú úroveň, aby už nebolo potrebné tvoriť pre každú zbierku rukopisov a tlačí samostatné modely. Pre používateľov by malo ísť o akúsi „čiernu skrinku“ (black box), v ktorej umelá inteligencia sama vyberie z integrovaných modelov najvhodnejší model transkripcie historických tlačí, rukopisov, strojopisov a iných dokumentov, ktoré používateľ chce študovať alebo sprístupniť. K tomuto cieľu však vedie dlhá cesta a nevyhnutnosť tvorby množstva parciálnych modelov.

Považujem za dôležité, aby súčasťou spoločného medzinárodného úsilia boli aj slovenskí a českí odborníci a aby budúca „čierna skrinka“ bola pripravená poskytnúť pomoc všetkým pri transkripcii historických zbierok a dokumentov.

V súčasnej fáze vývoja je dôležité zamerať pozornosť na prípravu parciálnych modelov transkripcie rukopisov a historických tlačí, a to na základe väčších zbierok, ktoré obsahujú stovky a tisíce strán[9]. Odporúčame zamerať sa na dokumenty v západoslovanských jazykoch, češtine, slovenčine, hornolužickej a dolnolužickej srbčine a poľštine. Charakter zbierok si vyžaduje aj pozornosť latinčine, nemčine a maďarčine. Mali by sme na základe vlastných modelov vytvárať jeden integrovaný model pre rukopisné dokumenty a jeden pre staré a vzácne tlače. To je úloha, ktorú za nás nikto neurobí.

Súčasný stav výskumu a aplikácií

Existujúce informačné zdroje k téme OCR sa, na jednej strane, týkajú pokračujúcich teoretických výskumov zameraných na samotnú umelú inteligenciu. Autormi teoretických diel sú najmä *informatici a matematici*. Na druhej strane sú diela, ktorých autori sú

[8] Výskum bol predtým financovaný ako súčasť projektu *tranScriptorium.* Tento projekt získal finančné prostriedky zo siedmeho rámcového programu Európskej únie pre výskum a technologický rozvoj podľa dohody o grante č. 600707.

[9] V prípade záujmu o transkripciu jednotlivých kratších dokumentov je možné skúsiť použitie niektorého z verejne dostupných modelov transkripcie s podobným typom písma, tlače alebo rukopisu.

z prostredia spoločenských a humanitných vied a odborov, teda *digital humanities*. Tí sa venujú téme OCR a HTR[10] z *používateľského hľadiska*, tj. z hľadiska praktickej použiteľnosti existujúcich nástrojov a platforiem OCR. Okrem toho sa teoretické alebo používateľské príspevky dajú rozdeliť do dvoch skupín podľa toho, či sa venujú OCR *tlačených* alebo *rukopisných* diel (HTR).

Komplexný prehľad projektu READ obsahuje projektová štúdia (Mühlberger 2016) a kolektívna štúdia výskumníkov READ (Mühlberger et al. 2019), ktorá je prvým publikovaným prehľadom o tom, ako je softvér HTR+[11] využívaný širokou komunitou odborníkov a ktorá ukazuje súčasnú aplikáciu technológie rozpoznávania rukopisov v sektore kultúrneho dedičstva. Táto štúdia popisuje aj vývoj metód rozpoznávania znakov.

Od polovice 20. storočia sa rozpoznávanie znakov tlačených a rukopisných dokumentov rozvíjalo spoločne s OCR. Najprv sa naskenované obrázky tlačeného textu konvertovali na strojový kód a porovnávali sa s hotovými šablónami písma. Tlačené dokumenty obsahujú znaky z vopred definovanými, hotovými znakovými súbormi, a preto je porovnávanie jednoduchšie. Avšak, aj softvéry OCR pre tlačené znaky sú schopné ďalšieho „doučovania“.

V porovnaní s tlačenými textami však rukopisné texty predstavujú odlišný problém kvôli množstvu odlišností rukopisov, rúk, zmien rukopisov v čase, množstvu glyfov, tokenov, osobných a jazykových štýlov ap. Rukopisy sa stali novou výzvou pre informatikov. Najprv, v 80. rokoch 20. storočia, sa výskum a vývoj rozpoznávania rukopisov rozvíjal s používaním *štatistických* metód. V 90. rokoch nasledoval výskum a vývoj rozpoznávania vzorov v kombinácii s *umelou inteligenciou* a vývoj *hlbokých neurónových sietí* v rokoch 2000 a 2010. Išlo aj o obdobie významného rozvoja a zvyšovania kapacít informačných a komunikačných technológií.

Vo viacerých vyspelých krajinách sa realizovali projekty masovej digitalizácie a vznikli mohutné digitálne repozitáre a archívy tlačených a rukopisných dokumentov[12]. Po masovej digitalizácii nastal čas aj na využívanie digitálneho obsahu získaného digitalizáciou rukopisov. Ak sa má z naskenovaných obrazov rukopisných dokumentov získať použi-

[10] HTR – Handwritten Text Recognition

[11] HTR+ – Handwritten Text Recognition. Softvér HTR+ spoločnosti Transkribus zatiaľ nemôže okamžite zahájiť automatický prepis, ale najprv musí byť vyškolený na konkrétny typ písma a rukopisu.

[12] Na Slovensku išlo o mimoriadny a v európskom kontexte bezprecedentný národný projekt masovej digitalizácie a konzervovania v gescii Slovenskej národnej knižnice (SNK) v Martine s názvom *Digitálna knižnica a digitálny archív* 2012–2015. Jeho iniciátorom a autorom bol Dušan Katuščák (Katuščák et al. 2008, 2011a, 2011 b, 2011c, 2021 a i.). Projekt sa čiastočne realizoval na základe zmluvy medzi SNK a Úradom vlády SR zo 7. marca 2012 o poskytnutí nenávratného finančného príspevku vo výške vyše 49 miliónov eur. Vybudovaná je unikátna infraštruktúra: 20 skenerov, z toho 10 digitalizačných robotov a poloautomatov, archív na dlhodobú ochranu digitálneho obsahu, platforma Slovakiana na sprístupňovanie digitálnych dokumentov, vytvorených je 73 nových pracovných miest. Cieľom bolo digitalizovať ca tri milióny dokumentov a fakticky celý slovacikálny knižničný fond, knihy, noviny, časopisy, zborníky ai. Unikátnosť projektu spočívala v integrácii masovej priemyselnej digitalizácie a priemyselného konzervovania degradujúceho kyslého papiera. Po podstatných zmenách manažmentu v roku 2012 sa do roku 2021 sa digitalizovalo len ca 10 % z plánovaného objemu a celkove sa použilo v SNK ca 60 miliónov eur. Masová deacidifikácia papiera sa nerealizuje, takže papier ako nosič ďalej nevratne degraduje (nevratný termodynamický dej). Digitálne dokumenty nie sú dostupné online. Stav digitalizácie je čiastočne kriticky popísaný v analýzach Ministerstva kultúry Slovenskej republiky (MKSR, 2019 a MKSR, 2020).

teľný, editovateľný text, je možné použiť pokročilú technológiu rozpoznávania *Transkribus* – stroje *HTR+* a *PyLaia*[13].

Projekt má všetky atribúty metodológie *digital humanities*. K týmto atribútom patrí najmä: a) kooperácia bádateľov; b) scientizácia v spoločenských a humanitných odboroch; c) interdisciplinarita; d) tímovosť (medziinštitučná, medzištátna, univerzity, knižnice, archívy, galérie, múzeá); e) výrazné zapojenie informatikov do výskumu, vzdelávania a sprístupňovania poznatkov; f) umelá inteligencia (umelé neurónové siete, Hidden Markov Model – HMM).

Pokrok vo výskume

O pokroku v rozpoznávaní tlačeného textu založenom na optickom rozpoznávaní tlačeného písma píše (Hodel et al 2021). Hodel sa venuje aj najdôležitejšiemu praktickému aspektu transkripcie, totiž otázke, čo je presnosť či chybovosť transkripcie. Hodel na základe empirických údajov z výskumu READ a opierajúc sa o poznatky Güntera Mühlbergera (2019) uvádza tri triedy chybovosti.

Hodel považuje za potvrdené a overené konštatovanie, že: a) ak je hodnota chybovosti znakov CER[14] nižšia ako 10 %, čo je 10 a menej chýb na sto znakov, tak výsledok transkripcie je dobrý, čitateľný a, ak je to účelné, je možné ďalšie editovanie výstupu; b) ak je chybovosť znakov CER ≤ 5 %, tak výsledok transkripcie je veľmi dobrý; c) ak je chybovosť znakov CER pod 3 %, potom je možné považovať výsledky transkripcie za výborné a chybovosť znakov CER pod 2,5 % za excelentné.

Hodelovi ide o cieľ, transkripcia bez tréningu. Konštatuje, že na tvorbu optimálneho univerzálneho modelu transkripcie rukopisov rôznych rúk, štýlov, typov písma, období ap., ktorý by si už zakaždým nevyžadoval prípravu samostatných modelov, je nevyhnutné mať čo najväčšie množstvo excelentných modelov. Usudzuje, že tieto modely transkripcie by mali byť pravdepodobne vyvíjané pre rôzne podobné triedy rukopisov, napríklad kurrentské písmo 19. storočia, ktorý je práve predmetom jeho pozornosti.

Ku pokroku v oblasti optického rozpoznávania znakov (OCR) prispieva (Strobel et.al 2020). Na základe analýzy efektívnosti niektorých systémov OCR tlačených nemeckých historických novín (fraktúry) autori dospeli k záveru, že dostatočná tréningová vzorka (*ground truth*) je 50 novinových strán. Svoje zistenia opierajú o porovnania piatich sys-

---

[13] *PyLaia* je nástroj na rozpoznávanie rukopisného textu, ktorý je podporovaný okrem stroja CITlab-HTR+. Tieto dva stroje fungujú dosť podobne, a tak zvyčajne sú výsledky podobné v chybovosti znakov (CER). Jediným rozdielom je, že v *PyLaia* môžu používatelia sami nastaviť niekoľko parametrov. Zmeniť sa dá aj sieťová štruktúra *PyLaia* – čo je príležitosť pre ľudí, ktorí poznajú strojové učenie. Úpravy neurónovej siete je možné vykonať prostredníctvom úložiska Github. *HTR+* zvyčajne poskytne lepšie výsledky so zakrivenými alebo otočenými čiarami, ale je možné, že *PyLaia* bude v tomto čoskoro schopná držať krok. Ak by bolo potrebné použiť nástroj *Text to Image*, treba použiť *HTR+*. Pre *PyLaia* to však ešte nie je implementované. Dokumenty, ktoré boli transkribované pomocou modelu *PyLaia* je možné prehľadávať pomocou plnotextového vyhľadávania (Solr) v *Transkribuse*.

[14] CER (Character Error Rates) je miera chybovosti znakov (porovnáva pre danú stranu celkový počet znakov (n) vrátane medzier s minimálnym počtom vložení (i), nahradenia (s) a vymazania (d) znakov, ktoré sú potrebné. získať výsledok *Ground Truth*. Ide teda o chyby v porovnaní s presným textom. Vzorec na výpočet CER je nasledujúci: CER = [(i + s + d) / n]*100. Každá malá chyba v prepise je štatisticky plnohodnotná chyba. To znamená, že každá chýbajúca čiarka, „u“ namiesto „v“, dodatočná medzera alebo dokonca veľké písmeno namiesto malého písmena sú zahrnuté v CER ako chyba.

témov OCR: 1. ABBYY FineReader XIX10 (FRXIX) z roku 2005, 2. ABBYY FineReader Server 11 (FRS11) vložený v minulých verziách do systému 3. *Transkribus* a *Transkribus HTR+*, 4. Kraken, 5. Tesseract.

Drobac (2020) poskytuje pohľad na efektivitu OCR v historických novinách a časopisoch vydávaných vo Fínsku. Fínska národná knižnica vytvorila pomocou programu ABBYY FineReader pre historický text korpus OCR s viac ako 11 miliónmi strán. Odhadovaná presnosť textu OCR bola medzi 87 % – 92 % na úrovni znakov, čo je na vedecký výskum dosť málo.

Martinek et al. (2020) predstavuje vo svojej teoretickej experimentálnej štúdii systém segmentácie tlačeného textu a OCR. Zaoberá sa súborom metód, ktoré umožňujú vykonávať OCR historických tlačí v nemčine na základe malého množstva cvičných údajov. Popisuje svoj OCR systém, ktorý využíva rekurentné neurálne siete. Sústreďuje sa na parciálne procesy systému OCR, a to hlavne na analýzu rozloženia stránky, vrátane segmentácie textového bloku a riadkov, a na samotné OCR. Popísané experimenty sú zamerané na určenie najlepšieho spôsobu dosiahnutia dobrých výsledkov OCR pre historické nemecké tlačené dokumenty. Na experiment použili digitalizovaný archívny materiál z projektu *Porta fontium* z česko-bavorského pohraničia. Konkrétne išlo o 10 strán z novín Ascher Zeitung z druhej polovice 19. storočia tlačených fraktúrou. Na tréning použili sedem strán, na validáciu jednu stranu a na hodnotenie efektívnosti dve strany. Ďalších 15 strán použili na tréning identifikácie a segmentácie šablóny strany. Získané výsledky považujú autori za porovnateľné alebo dokonca lepšie ako výsledky niekoľkých najnovších systémov, napríklad *Transkribus*. Pri fraktúre z nemeckých novín dosiahli v porovnaní s inými systémami tieto hodnoty CER: Porta fontium CER 0,024 %; Tesseract (deu_frak) CER 0,053 %; Tesseract (Fraktur) CER 0,045 %; *Transkribus* CER 0,027 %.

Téme rozpoznávania novovekých tlačených textov písaných fraktúrou sa venuje Martin Kišš (2018) vo svojej diplomovej práci. Svoj výskum založil na nástroji *TensorFlow*, ktorý pôvodne vyvinula spoločnosť Google a je k dispozícii ako *open source* platforma pre strojové učenie. Súčasťou jeho prístupu je vstavaný generátor umelých historických textov. Pomocou tohto generátora vytvoril umelú dátovú sadu, na ktorej trénoval neurónovú sieť na rozpoznávanie riadkov. Túto neurónovú sieť otestoval na reálnych historických riadkoch textu a dosiahol po natrénovaní úspešnosť 89,0 % presnosti znakov.

Význam a vlastnosti platformy Transkribus

Vytvorenie výskumnej platformy *Transkribus* bolo okrem základného výskumu jedným z hlavných cieľov projektu READ. Približne 2,5 milióna EUR z 8,2 milióna EUR sa investovalo do rozvoja tejto výskumnej infraštruktúry. Teraz vznikajú nadväzujúce projekty, v ktorých pokračuje základný aj aplikovaný výskum. Osvojovanie si platformy *Transkribus* môže mať aj významné ekonomické efekty.

Podľa údajov z internej dokumentácie projektu READ sa trhové ceny manuálneho prepisu historických rukopisov pohybujú od 10 € až do 30 € alebo aj viac za jednoduchú angličtinu, nemčinu, latinčinu za konkrétny rukopis. Ak predpokladáme 15 € za stranu ako priemerné náklady, tak v projekte READ výskumníci generovali peňažnú hodnotu 4–6 miliónov EUR. Tieto údaje sú pridanou hodnotou a potenciálnym zdrojom rozvoja novozaloženého združenia READ-COOP[15] a presvedčivým potvrdením základnej kon-

[15] READ-COOP. [cit 1. 10. 2022] Dostupné z: O nás – READ-COOP (*readcoop.eu*). V októbri roku 2022 malo združenie 113 členov z 27 krajín. Jedinou členskou krajinou zo strednej a východnej Európy bolo v tom čase Slovensko.

Tab. 1 Ceny automatickej transkripcie

Suma / kredity[1]	**Stroj PyLaia**	**Stroj PyLaia 2**	**Stroj HTR+**	**Stroj HTR+3**
	Počet strán rukopis/cena €	*Počet strán tlač/cena €*	*Počet strán rukopis/cena €*	*Počet strán rukopis/cena €*
648 €/3000	3 000/0,216	18 000/0,036	2 400/0,27	15 000/0,043
1944 €/10000	10 000/0,194	60 000/0,0324	8 000/0,24	50 000/0,038

cepcie výskumu smerujúcej k novým poznatkom a súčasne ku komerčnému využitiu nástrojov, ktoré sú výsledkami aplikácie nových poznatkov. Orientačné náklady na transkripciu vrátane DPH sú v nasledujúcej tabuľke[16] č. 1.

Predstavitelia *digital humanities* na Slovensku majú k tejto iniciatíve rozličné postoje. Od nadšených prejavov súhlasu a obdivu až po veľmi rezervované až odmietavé postoje (typu „to nie je nič pre nás“, „máme iné starosti“, „umelá inteligencia nenahradí nás expertov“). Často ide o reakcie, ktoré na jednej strane síce verbálne deklarujú záujem o „digitalizáciu“ a „umelú inteligenciu“, no na druhej strane svedčia o nepochopení a nedostatočných vedomostiach o problematike a možnostiach digitalizácie a využitia umelej inteligencie. Postoje svedčia skôr o uprednostnení tradičných paradigiem práce a výskumu než o reálnej snahe hľadať inovatívne nástroje sprístupnenia a interpretácie nášho obrovského historického písomného dedičstva ako súčasti európskeho kultúrneho dedičstva.

Pokiaľ ide o transkripciu *slovenčiny*, tá sa ocitla v zozname jazykov v záverečnej správe o projekte READ vďaka našej iniciatívnej práci, a to bez akejkoľvek podpory a v podstate bez záujmu národných inštitúcií, archívov, knižníc, múzeí a akademického sektora. Išlo o prácu, ktorej sme venovali od roku 2017 najmenej 3000 hodín a ktorú autor tohto príspevku financoval do roku 2020 len z vlastných zdrojov. Dosiahnuté výsledky, know-how a skúsenosti nás viedli k úsiliu zaviesť revolučnú a inovatívnu platformu *Transkribus*[17] na Slovensku a v Česku[18], najmä do systému vzdelávania, ako aj do praxe pamäťových a fondových inštitúcií prostredníctvom projektov výskumu a vývoja. Samozrejme, rešpektujeme pritom aj iné nástroje transkripcie.

Platforma *Transkribus* je slobodný softvér (*open source*) s garanciou bezpečného používania pre registrovaných klientov platformy. Svoj účet si môže vytvoriť každý a potom si môže zadarmo stiahnuť *Transkribus Expert Client* alebo môže používať jednoduchší nástroj *Transkibus Lite*. Na pripojenie počítačov alebo mobilných zariadení klientov k platforme je k dispozícii rozhranie API. Väčšinu softvérových nástrojov tvoria slobodné softvéry, ktoré je možné získať z *GitHubu*.

[16] Manuálna transkripcia: 10–15 €/strana; automatická transkripcia – Transkribus: ca 0,12 € – 0,14 €/strana. Prepočet podľa: Transkribus Credits & Pricing – READ-COOP (*readcoop.eu*).

[17] V roku 2017 autor pracoval s verziou *Transkribus Expert Client* v1. 3. 7. V októbri roku 2022 bola k dispozícii verzia 1. 22. 0.

[18] HITEXT. Slezská univerzita v Opave pripravila v r. 2022 návrh projektu aplikovaného výskumu s akronymom HITEXT v programe NAKI III. Projekt sa v r. 2022 posudzuje. Mimo toho problematiku riešime v rámci vzdelávania a v projekte študentskej grantovej súťaže v r. 2022.

Alternatívy platformy Transkribus

V štúdii sa venujeme výlučne platforme *Transkribus* a transkripcii rukopisných zbierok a okrajovo aj transkripcii tlačí. Existuje však celý rad iných nástrojov transkripcie. Napríklad *OCR4all*, ktorý bol vyvinutý na digitalizáciu starých tlačí. Ďalej aplikácia *eScript*, ktorá slúži na transkripciu rukopisov a tlačí. Nástroj *Rescribe* je určený pre stolné počítače na OCR na obrazových súboroch, súboroch PDF a Knihách Google. Jedným z použiteľných nástrojov transkripcie je aj *Pero.cz*. Systém *ABBYY Cloud OCR SDK* je veľmi kvalitná aplikácia v cloude prostredníctvom webového rozhrania API. Aj ku *ABBYY Cloud OCR SDK* existuje viac ako 10 alternatív. Najlepšou alternatívou je *Online OCR*, ktoré je zadarmo. Ďalšie skvelé stránky a aplikácie podobné *ABBYY Cloud OCR SDK* sú aj *Kofax Omnipage, Geekersoft OCR Word Recognition a i2OCR*. K dispozícii je aj komerčný *Quartex* (Adam Matthew Digital 2018). Pred výskumníkmi v budúcnosti stojí úloha vypracovať metaanalýzu s kritériami hodnotenia funkcionality a kvality nástrojov, aplikácií a platforiem transkripcie. Predmetom tejto štúdie však nie je hodnotenie iných systémov transkripcie.

READ-COOP

Projekt READ skončil 30. 6. 2019. Následne vzniklo medzinárodné združenie READ-COOP SCE (*Societas Cooperativa Europeae* – SCE), a to 1. 7. 2019. Jeho cieľom je udržať a ďalej rozvíjať platformu *Transkribus*. Odborníci a inštitúcie majú záujem o pokračovanie a vývoj služby *Transkribus*. V súčasnosti, v októbri roku 2022, je 27 členov a viac ako 90 000 používateľov *Transkribus*, ktorí pracujú s touto platformou.

Obr. 1 Rozšírenie platformy Transkribus v Európe (Zdroj: readcoop.eu, podľa stavu v septembri 2022. Aktuálne august 2022: Members of READ-COOP SCE – READ-COOP (readcoop.eu))

Projekt SKRIPTOR[19]

Slovenskí odborníci reagujú na nové trendy OCR a výskumu historických dokumentov projektom SKRIPTOR (Katuščák a Nagy et al., 2019). Projekt má európsky aj národný rozmer.

Projekt SKRIPTOR priamo nadväzuje na európsky projekt READ. Technologické a vedecké inovácie projektu READ sú založené na využívaní umelej inteligencie a metodológie digitálnych humanitných vied. Úlohou výskumníkov projektu SKRIPTOR je implementácia a rozšírenie najnovších technologických inovácií a poznatkov o efektívnom prístupe odbornej a laickej verejnosti k slovenskému i zahraničnému písomnému dedičstvu.

Strategickým cieľom projektu SKRIPTOR je vytvárať podmienky na vnútroštátnej úrovni pre kompetentné partnerstvo slovenských výskumných pracovníkov s popredným európskym výskumom, nadviazať a potom sa aktívne zapojiť do mnohostrannej vedeckej európskej spolupráce. Projekt SKRIPTOR sa realizuje v oblasti histórie a archívnictva. Presahuje tiež do knižničnej a informačnej vedy.

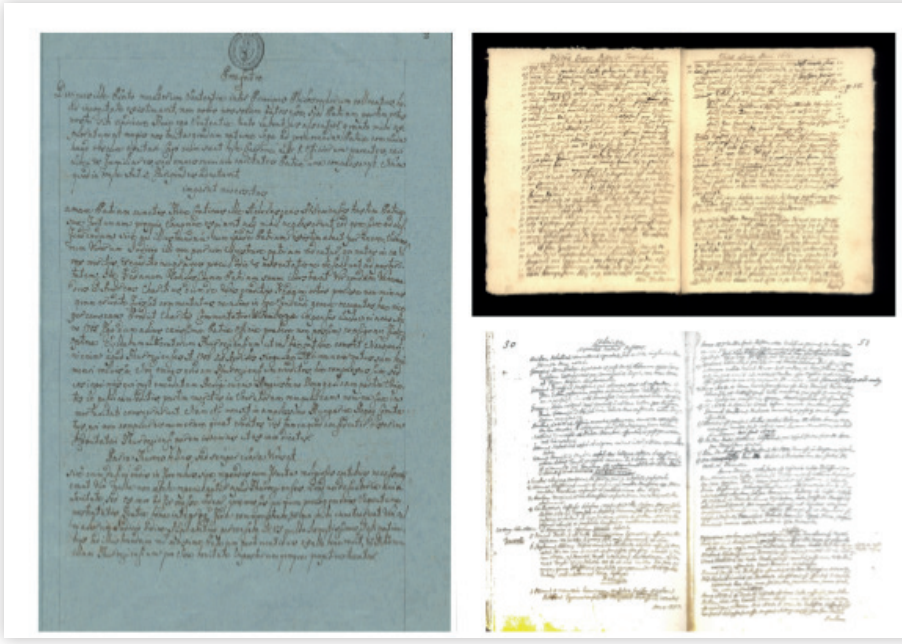
Projekt SKRIPTOR je zameraný na dokumenty novoveku. Zbierky, ktoré sú predmetom preskúmania a sprístupnenia, môžu však zahŕňať aj významné texty novších dokumentov a inkunábul, tlačené materiály zo 16. storočia, historické časopisy, noviny, ako aj cenné materiály z 18. až 20. storočia.

Cieľom tvorby nových modelov s použitím platformy *Transkribus* je potvrdiť jej efektívnosť a dosiahnuť pri našich zbierkach zníženie ceny transkripcie z 30 € za manuálnu transkripciu strany na menej ako 1 €/strana za automatickú transkripciu textov.

V projekte SKRIPTOR sme predbežne zvolili na výskum a experimentálnu transkripciu tieto zbierky: 1. Slovenská a česká fraktúra (švabach i antikva); 2. Andrej Kmeť – osobná rukopisná korešpondencia; 3. Martin Lauček – Collectanea; 4. Postila Izáka Abrahamidesa Hrochotského z rokov 1600–1601; 5. Postila Juraja Schmidelia-Kováčika z rokov 1598–1607; 6. Kanonické vizitácie Banskobystrickej diecézy z 18.–19. storočia; 7. Hurban, J. M., rukopisné dokumenty; 8. rímskokatolícke matriky; 9. urbáre tereziánskej urbárskej regulácie; 10. parcelačné protokoly stabilného katastra; 11. kongregačné zápisnice, sedriálne protokoly; 12. ďalšie zbierky písomností identifikované počas archívneho výskumu.

Zatiaľ, v roku 2022, sú v projekte SKRIPTOR dostupné niektoré výstupy a súvisiace aktivity: Publikácie: NAGY, I. (2021), TOMEČEK, O. (2021), BÔBOVÁ, M. (2021), KATRENIAK, M. (2022), KATUŠČÁK, D. (2020, 2021), KOVÁČOVÁ, K. (2022). Ďalej návrh projektu HITEXT v Česku TAČR (2020) a NAKI III (2022): KATUŠČÁK, D. (2020 a 2022). Účasť na študentskej vedeckej konferencii v Opave, aktivity v študentskej grantovej súťaži SGS/5/2022 (SGS SU Opava). Dôležité je osvojenie si funkcionality platformy *Transkribus* a prenos poznatkov do procesu vzdelávania na Slovensku a v Česku.

[19] Projekt Agentúry na podporu vedy a výskumu – APVV-19-NEWPROJECT-17816 (2020–2024). Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov [Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts].



Obr. 2 Rukopis Martina Laučeka. Od krasopisu k voľnejšiemu rukopisu

Pracovný postup transkripcie

Na základe vlastných skúseností chápeme transkripciu ako komplexný proces, ktorý predpokladá najmä odhodlanie, dostupnosť finančných zdrojov a infraštruktúry. Hlavné procesy sú:

Príprava. Najmä: Informačný archívny prieskum (heuristika), identifikácia možných zbierok a dokumentov, vyriešenie podmienok dostupnosti zbierok a dokumentov, kvantifikácia a výber dokumentov na transkripciu (počet strán a homogénnosť rukopisov), dohoda s vlastníkom alebo správcom zbierky o mieste a spôsobe snímania a o právach.

Snímanie. Najmä: skenovanie, fotografovanie dokumentov, pomenovania a organizácia adresárov a súborov v počítači, archivovanie zdrojových súborov (TIFF, RAW) a zálohovanie derivovaných súborov (JPG, PDF, PNG ai.).

Inštalácia Transkribus Expert Client a práca s platformou *Transkribus*. Najmä: zoznámenie sa s dokumentáciou *Transkribus*, voľba formátu obrázkov pre *Transkribus*, kontrola kvality a príprava obrázkov na nahrávanie do *Transkribusu*, voľba spôsobu nahrávania súborov, vytvorenie vlastnej zbierky, nahrávanie zvolených súborov do platformy *Transkribus* do zbierky.

Manuálna transkripcia. Najmä: výber vzoriek strán na manuálnu transkripciu podľa špecifík rukopisu, rozhodnutie o zdieľaní zbierky so spolupracovníkmi a o ich úlohe, manuálna transkripcia vzorky pre cvičný súbor.

Obr. 3 Študentka knihovníctva Slezskej univerzity sníma v archíve v Jeseníku so ScanTent a DocScan rukopisný text pre svoju záverečnú prácu

Segmentácia strán a metadáta v *Transkribus Expert Client.* Najmä: segmentácia strán alebo celých súborov, kontrola kvality a oprava manuálnej transkripcie a segmentácie, metadáta dokumentu, metadáta stránky, štrukturálne metadáta, komentáre, KWS[20].

Tvorba modelu transkripcie v *Transkribus Expert Client*. Najmä: učenie stroja pre model transkripcie, kontrola kvality a efektívnosti modelu a korekcie cvičného súboru, opätovné spustenie tvorby modelu a kontrola kvality modelu, voľba stránok v kvalite *ground truth*, použitie modelu na transkripciu všetkých segmentovaných strán v zbierke.

Sprístupnenie a použitie výsledkov transkripcie. Najmä: export výsledkov rôznymi spôsobmi a v rôznych formátoch, editovanie a korekcie výsledkov transkripcie v *Transkribus Lite*, použitie modelu transkripcie, sprístupnenie výsledkov transkripcie v lokálnej sieti alebo zverejnenie výsledkov transkripcie online na využívanie na internete cez *read & search* (viď ďalej).

Experiment so zbierkou listov Andreja Kmeťa

O automatickej transkripcii rukopisných textov už desiatky rokov snívajú historici, lingvisti, archivári, knihovníci, dokumentaristi a všetci ďalší, ktorí prichádzajú do styku s rukopisnými textami.[21] Postupne sa automatický prepis rukopisov stáva skutočnosťou. Je

[20] KWS (The Keyword Spotting) je výkonný nástroj na vyhľadávanie, ktorý pomáha vyhľadať podobné obrazy slov v dokumentoch. Hlavnou výhodou je, že nie je potrebné, aby sa dokumenty definitívne transkribovali. Jednoducho spustí nejaký model transkripcie textu a potom je okamžite možné prehľadávať dokumenty. KWS spoľahlivo nájde slová a frázy (varianty obrazov textu). Tento nástroj ukáže, na ktorých stránkach bolo nájdené zadané kľúčové slovo, a zobrazí úryvok ukážky. Okrem toho poskytne obrázok medzi hodnotami 0 a 1 (0 = najnižšia a 1 = najvyššia), aby sa zhodnotila kvalita výsledkov hľadania.

[21] Pamätám si, koľko úsilia a času musel v minulosti vynaložiť Pavol Vongrej na prepis 20 400 veršov rukopisného diela *Matora* Michala Miloslava Hodžu, či Viliam Sokolík na prepis časti korešpondencie medzi A. Kmeťom a V. Riznerom. V roku 1991 v spolupráci s Ing. Jánom Mišíkom skúsil som použiť systém rozpoznávania znakov na automatický prepis ručne písaných katalogizačných záznamov zo starého katalógu Slovenskej národnej knižnice (Matice slovenskej). V dôsledku toho bola účinnosť

za tým mohutný medzinárodný základný výskum v oblasti umelej inteligencie a tisíce hodín práce.

Transkribus, pochopiteľne, nenahrádza odbornú a vedeckú erudíciu historikov a archivárov. Preto je pochopiteľný aj ich rezervovaný postoj. Umelá inteligencia nesúťaží s odborníkmi. Pomáha im. Automatická transkripcia môže byť len jedným z krokov vedeckej práce odborníkov. Ďalej nasleduje historický výskum textu a kontextu transkribovaných textov a informácií, editovanie textov získaných transkripciou, identifikácia entít, tvorba kľúčových slov, metadát, ktoré sú v texte objavené (dátumy, mená osôb, názvy geografických jednotiek, korporácií a pod.).

Zmyslom rozsiahlejšej transkripcie s použitím špičkovej platformy *Transkribus* je uľahčenie čítania a sprístupnenie unikátnych zbierok, dokumentov, archívnych jednotiek, ktoré sa nachádzajú v archívoch spravidla len v jednom exemplári. V tom je rozdiel medzi výskytom jednotiek v knižniciach a archívoch. V archívoch sú jedinečné, autentické originálne dokumenty, zbierky, archívne jednotky, kým v knižniciach sú tituly dokumentov, ktoré majú často stovky až tisíce exemplárov. Unikátne archiválie je potrebné sprístupniť. Cesta ku sprístupneniu vedie cez ich transkripciu.

Po transkripcii historických textov a rukopisov je možné digitálny obsah editovať, interpretovať, použiť a sprístupniť na využitie v širšom meradle aj vo verejných informačných systémoch a službách. Navyše, transkribovaný originálny text, napríklad v latinčine, maďarčine, nemčine alebo v inom jazyku, je možné aspoň približne ďalej automaticky preložiť do iného jazyka. Tým sa dosť podstatne mení charakter práce archivárov a historikov. Výsledkom mojej práce sú modely transkripcie rôznej kvality. Prehľad modelov obsahuje tabuľka.

Vysvetlivky k tabuľke:

Dátum: Dátum vytvorenia modelu (RRRRMMDD).
Metóda: Zvolená metóda transkripcie rukopisu (HTR+).
ID: Identifikačné číslo modelu v našich zbierkach a medzi všetkými modelmi *Transkribus* na vzdialenom serveri.

Tréningový súbor: Počet strán a počet riadkov, ktoré boli manuálne prepísané a použité na učenie (tréning) stroja v platforme *Transkribus*. Spolu sme na cvičenie postupne prepísali 211 strán. Z nich 185 slúži na tréning a 26 na validáciu (overenie). Základný prepis obsahoval 50 strán. S ním sme urobili prvý model. Z výsledkov transkripcie sme pridávali do tréningového modelu editované ďalšie strany a tvorili sme ďalšie modely.

Validačný súbor: Počet strán a riadkov, ktoré sme zvolili z celkového počtu prepísaných strán na overenie presnosti učenia.

Presnosť CER: Percentuálne vyjadrenie chybovosti znakov vo vstupnom súbore a vo validačnom súbore. Pri rukopisoch je prakticky vylúčené, aby manuálny prepis bol 0,0 %.

transkripcie IRIS OCR približne 35/40 % a transkripcia bola nepoužiteľná. Signálnu informáciu o práci s platformou Transkribus som zverejnil v roku 2018 v jednom blogu a v statuse na Facebooku. Bol som prekvapený deklarovaným záujmom o túto prácu. Je to pochopiteľné, pretože mnohí historici, jazykovedci, knihovníci, pedagógovia sú čoraz vzdelanejší v používaní nových technológií vo svojej práci a chápu, že inovácie, ktoré im prácu uľahčia, sú dôležité.

Tab. 2 Prehľad experimentov s modelmi transkripcie rukopisnej korešpondencie Andreja Kmeťa

Dátum	Metóda	ID modelu	Tréningový súbor		Validačný súbor		CER		Cykly	CER/WER	
			strany	riadky	strany	riadky	tréningový	validačný	(epochy)	znaky	slová
20190125	CITlabHTR+	10135	125	22549	26	3497	1,15 %	3,37 %	200	5,97 %	21,60 %
20190201	CITlabHTR+	10410	152	29905	46	4499	1,27 %	2,97 %	200	6,19 %	22,13 %
20190205	CITlabHTR+	10548	166	29411	46	4573	1,37 %	1,84 %	200	5,91 %	21,87 %
20201012	CITlabHTR+	26809	111	18071	98	2921	0,44 %	7,25 %	500	6,08 %	21,87 %
20210410	CITlabHTR+	31888	119	19291	13	3126	1,15 %	5,16 %	200	3,77 %	12,27 %
20210821	CITlabHTR+	**36009**	185	28672	26	4703	**1,87 %**	**5,79 %**	200	2,48 %	7,73 %

Počet cyklov: Počet cyklov, tzv. epoch, ktoré stroj použil na učenie (tréning).
CER/WER: Hodnoty vyjadrujú skutočnú praktickú, používateľskú presnosť resp. chybovosť znakov CER a chybovosť slov WER[22] v šiestich modeloch z rokov 2019–2021, ktoré sú vo vlastníctve autora. Všetky modely sme testovali na jednej, čo najpresnejšie pripravenej dvojstránke v kvalite FINAL v zbierke ID 115514. Ide o list Andreja Kmeťa Ľ. V. Riznerovi (dokument ID 621673).
Priemer prepočítanej chybovosti znakov v šiestich modeloch je CER 5,0 %, pričom päť z nich sme vytvorili na cvičných súboroch rôznej kvality, hlavne v statuse *In Progress*. Na praktickú transkripciu ďalších stoviek strán bude však najvhodnejšie použiť model 36009. Na cvičenie tohto modelu sme použili kvalitne pripravené strany – v kvalite *ground truth*. Z hľadiska presnosti transkripcie ďalších listov Andreja Kmeťa považujeme výsledky Modelu 36009 s hodnotami CER 2,48 % a WER 7,73 % za najlepšie.
Uvedené údaje v stĺpci CER/WER nevyjadrujú presnosť transkripcie pri vytváraní modelu s vopred pripravenými súbormi na tréning (1,87 %) a validáciu (5,79 %), ale najlepšie hodnoty, ktoré sa týkajú jednotlivých *strán*. Preto tie hodnoty sú odlišné. CER/WER 2,48 % a 7,73 % sú len najlepšie hodnoty, ktoré sa týkajú v danom modeli jednej *strany*, ktorú zo zbierky treba vybrať náhodne a ktorá nie je vopred nijako transkribovaná. Samotná hodnota WER nemá nejaký praktický zmysel, pretože ak použijeme v *Transkribuse Tools/Compare text version,* zistíme, že napr. interpunkcia, dĺžeň, mäkčeň, bodka… v slove má dištinktívnu rolu, a ak je navyše, alebo chýba v transkribovanom texte v porovnaní s GT (*Ground Truth*), tak stroj slovo považuje za chybné, hoci pre používateľa je text jednoznačne zrozumiteľný a nesťažuje jeho použitie. Hodnoty WER sa používajú väčšinou v matematickej lingvistike, napr. v strojovom preklade.

Výsledky transkripcie dokumentov priebežne organizujeme a zverejňujeme aj na internete cez nástroj, ktorý vyvinul tým READ-COOP a ktorý sa volá *read & search*. Verejný prístup k dokumentom je teda cez stránku *read & search – https://Transkribus.eu/r/slovakia-state/#/,* ktorej interfejs sme preložili do slovenčiny.

Všetky modely uvedené v tabuľke sme kvôli porovnaniu testovali na jednej, čo najpresnejšie pripravenej dvojstránke v kvalite FINAL v zbierke ID 115514. Ide o list Andreja Kmeťa Ľ. V. Riznerovi (dokument ID 621673). Chybovosť *slov* je *de facto* irelevantná, pretože chybný znak (napr. interpunkcia) spôsobuje vo väčšine prípadov aj chybovosť slova.

Priemer prepočítanej chybovosti *znakov* v šiestich modeloch je CER 5,0 %, pričom päť z nich sme vytvorili na cvičných súboroch a stranách rôznej kvality, ktoré boli hlavne v statuse *In Progress*.[23] Na praktickú transkripciu ďalších stoviek strán bude však najvhodnejšie použiť model 36009, ktorý sme vytvorili zo 185 strán cvičného súboru a 26 strán validačného súboru. Ukazuje sa, že najnižšie hodnoty presnosti CER vo validačnom súbore neznamenajú, že modely, ktoré sú v tabuľke na prvých piatich riadkoch v šiestom stĺpci a nie sú vytvorené na stranách *ground truth*, sú najvhodnejšie pre ďalšiu transkripciu.

Na poslednú prípravu tohto modelu sme použili kvalitne pripravené strany v kvalite *ground truth*. Z hľadiska presnosti transkripcie ďalších listov Andreja Kmeťa považujeme

[22] WER – *Word Error Rates*

[23] Statusy transkripcie sú: *New* (nový – stav pre novonahraté dokumenty), *In: Progress* (prebieha – automatická zmena stavu po úprave strany), *Done* (hotovo – stránka je prepísaná), *Final* (finálna verzia – stránka prepísaná a skontrolovaná), *Ground Truth* („základná pravda“ – 100 % správne prepísaná strana). Znamená to, že sa zaznamenáva práca s každou jednotlivou stranou a verzii strany sa môžu priradiť rôzne stavy v závislosti od toho, aký pokrok sa na nich už dosiahol.

výsledky Modelu 36009 s hodnotami CER 2,48 % a WER 7,73 % za najlepšie. V budúcnosti, na základe ďalších skúseností, zvážime poskytnutie tohto nášho modelu na voľné použitie pre podobné rukopisné zbierky.

Výber zbierky

Na experiment sme zvolili zbierku rukopisnej, prevažne slovenskej korešpondencie Andreja Kmeťa, uloženú v knižnici Slovenského národného múzea v Martine, a to po predchádzajúcom láskavom súhlase riaditeľky múzea. Niekoľko listov je v latinčine, maďarčine a časti listov aj v nemčine a češtine. Ide o listy Andreja Kmeťa z rokov 1841–1908. V oblasti vedeckého prístupu ku korešpondencii učencov v novoveku v duchu metodológie *digital humanities* je nepochybne najkomplexnejším zdrojom poznatkov medzinárodný výskum, ktorý inicioval a viedol Howard Hotson v rokoch 2014–2018 (Hotson 2019). V tejto štúdii nás korešpondencia zaujíma len ako rozsiahly rukopisný materiál, ktorý je vhodný na experimenty s automatickou transkripciou.

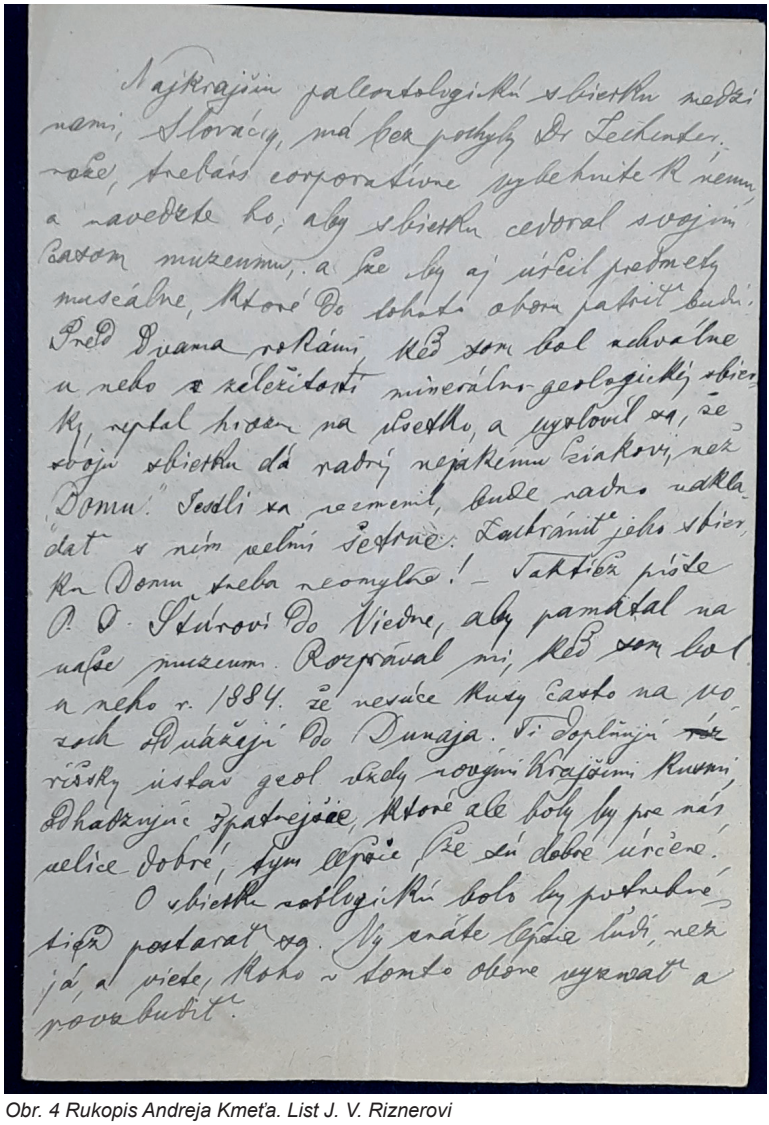
Osobnosťou Andreja Kmeťa, vrátane spracovania častí jeho korešpondencie, sa zaoberá systematicky Karol Hollý; a uvádza aj ďalšie zdroje, ktoré sa týkajú Kmeťovej rukopisnej pozostalosti (Hollý 2013, 2019).

Snímanie

Snímanie, teda skenovanie, presnejšie fotografovanie, prebehlo 23. – 30. 5. 2018 v Knižnici Slovenského národného múzea v Martine. Na snímanie sme použili zariadenie *ScanTent* (skenovací stan) a voľne dostupnú aplikáciu DocScan. *ScanTent* sme použili zámerne, aby sme overili celý navrhovaný pracovný postup *Transkribus*. Je známe, že mnohé archívy už majú časti zbierok viac-menej kvalitne skenované. Nami zvolené zariadenia majú význam v prípadoch, ak zbierky ešte nie sú skenované. Je známe, že zo študovní archívov bežní vedci a používatelia nesmú vynášať archiválie. Amatérske fotografovanie strán smartfónmi alebo fotoaparátmi je problematické, ak ide o väčšie súbory (tisíce strán). Preto je *ScanTent* a *DocScan* prijateľnou a dostupnou voľbou, ktorá je s určitými praktickými výhradami (formát, zaostrovanie, kvalita) prijateľná. Treba si však uvedomiť, že v tomto prípade ide o fotografovanie a nie o skenovanie v pravom technologickom zmysle slova. V budúcnosti by sme rozhodne použili na snímanie profesionálny skener a skenovanie v najvyššie dosiahnuteľnej kvalite (300–600 dpi).

Snímali sme kompletný obsah piatich krabíc. Niektoré listy boli na viacerých stranách, vyskytujú sa neúplné strany, vakáty a pod. Jeden obraz mohol obsahovať aj viac strán rukopisu. Vo fáze snímania sa vytvárajú obrazy a nie strany, pokiaľ sa strany nesnímajú oddelene. Niekedy je vhodnejšie listy snímať podľa strán, jednotlivo, pretože, ak sa sníma list ako dvojstrana, musí sa niekedy prácne usporadúvať poradie strán v následnom spracovaní obrazu, tzv. post processingu. V ďalšom kroku segmentácie textu je však možné jednotlivé strany ako bloky textu usporiadať do správneho poradia. Jednotlivé strany v listoch Andreja Kmeťa nenasledovali za sebou, takže na skenovanom obraze bola napríklad strana 3 a 1, na ďalšej 2 a 4.

Čas snímania asi 3000 strán bol spolu ca 15–20 hodín. Snímanie bolo v manuálnom režime *single* podľa jednotlivých listov, nie *series*, teda nie s automatickým snímaním po obrátení strany, nakoľko rukopisný materiál je na samostatných listoch rôzneho formátu. Časť materiálu tvoria originály listov, časť fotokópie. Najmä originály listov sú často na krehkom papieri, ktorý by si vyžadoval konzervačné zásahy. Vizitky a podobné menšie formáty papiera – softvér *DocScan* žiadal „priblížiť“, čo sme riešili podložením čistej



Obr. 4 Rukopis Andreja Kmeťa. List J. V. Riznerovi

stránky formátu A4 pod chýbajúce časti listu. Niektoré listy boli poškodené (chýbal roh, poškodené strany listu). Systém v takom prípade hlásil *no page found*. Riešili sme to použitím bielej strany ako podložky pod list, aj pod chýbajúce časti, potom *DocScan* zaostril.

Niektoré zložky sme museli snímať znovu, nakoľko sme nevenovali spočiatku potrebnú pozornosť zaostrovaniu. *DocScan* zaostruje na plochu listu na niekoľkých miestach. Zaostrenie indikujú červené a zelené značky. Keď je zaostrenie uspokojivé, zobrazí sa „OK“, potom možno stlačiť spúšť. Na snímanie sme použili mobilný telefón *Samsung Galaxy 6* s operačným programom *Android*, s ktorým vtedy fungoval *DocScan*. Nejasný bol pre nás spočiatku proces prenosu dát zo *Samsungu* (Android) do *MacBook Air* (operačný systém iOS). V súčasnosti je dostupný softvér *DocScan* aj pre operačný systém

iOS. Napokon sme použili počítač s Windows 10 a stiahli sme obrázky z Pictures zo Samsungu do iného počítača. Použitie systému *DocScan* a mobilného telefónu Samsung považujeme za vyslovene núdzové riešenie, pretože sme v ďalšej práci, najmä pri segmentovaní, zistili pomerne veľké množstvo neostrých častí strán. Keďže boli časti strany neostré, segmentácia bola nepresná a následne nebola ani transkribovaná. V budúcnosti by sme odporúčali používať pri rozsiahlych cenných zbierkach kvalitné profesionálne skenery a samotné skenovanie v najvyššej dosiahnuteľnej kvalite.

Systém *DocScan* je možné pri snímaní napojiť priamo na server a platformu *Transkribus* (v Innsbrucku či Rostocku), snímať a priamo zo snímania prenášať obrazy do platformy *Transkribus*. Túto možnosť sme nevyužili. Považovali sme za potrebné preveriť správnosť a kvalitu snímania. Niektoré operácie s *Transkribus* si vyžadovali použitie *Preview*, *Adobe Acrobat Pro DC verzia 2021.001*, *FileZilla Client verzia 3.61.0*, *ABBYY FineReader PDF 15*, *Zoner Photo Studio X* a i. Nástroje sme využili na úpravu orientácie textu, hromadné orezanie, konverzie formátov, vylúčenie duplicít, usporiadanie stránok v súbore, zlučovanie súborov ap.

Snímaný digitálny obsah (obrazy) bol: a) pripravený na ďalšie spracovanie v softvéri *DocScan* (identifikácia obsahu, metadáta), b) nahratý bez úprav na CD ROM na použitie u vlastníka zbierky podľa uváženia vedenia, c) obrazy boli pripravené na nahratie do platformy *Transkribus* a na ďalšie spracovanie v softvéri *Transkribus*. Nasledovalo nahrávanie na server *Transkribus*, segmentácia, tvorba modelov a transkripcia rukopisného textu.

Digitálny obsah sme rozdelili tak, ako sa nachádza v archívnych krabiciach. Napálili sme teda päť kompaktných diskov (CD), ktoré sme netranskribované protokolárne odovzdali vtedajšej riaditeľke etnografického múzea v Martine Dr. Márii Halmovej. Správcovia zbierky, archivári, teraz môžu použiť digitálny obsah a celý ho zverejniť. Ďalej môžu vložiť do každej krabice jeden kompaktný disk. Môžu rozhodovať tom, komu umožnia prístup k zbierke na disku alebo opäť umožnia prácu s pomerne krehkými papierovými originálnymi archívnymi listami. Transkribovaný obsah sprístupňujeme postupne cez softvér *read & search*, ktorý funguje ako „softvér ako služba“ (SaaS). Zatiaľ ešte len skúmame možnosti optimálnej prípravy metadát pre dokumenty a zbierky na zverejnenie cez *read & search.*

Nahrávanie súborov digitálnych obrazov

Snímané obrazy je možné spracovať buď lokálne, alebo ich upravovať po importe na vzdialený server *Transkribus*. Pred importom na server a pred používaním platformy *Transkribus* je potrebné zaregistrovať sa, stiahnuť si platformu *Transkribus Expert Client*. Pracovať je možné aj s nástrojom *Transkribus Lite*, v ktorom však nie je možné tvoriť vlastné modely transkripcie. Potom je potrebné vytvoriť si svoju vlastnú privátnu zbierku, ktorá je dostupná výlučne tomu, kto ju vytvoril, ak sa nerozhodne zdieľať ju s ďalšími používateľmi. Je možné, aby „prepisovač“, teda transkriber, umožnil prístup k niektorým operáciám napríklad študentom, operátorom, kooperantom. Môže umožniť prístup k vlastnej zbierke na prípravu cvičnej vzorky, editovanie po transkripcii a pod. Automatická transkripcia sa vykonáva výlučne na vzdialenom serveri s použitím infraštruktúry *Transkribus Expert Client*. Lokálne je možné s vlastnými dokumentami a zbierkami pracovať podľa potreby.

Pred importom súborov je potrebné vytvoriť si vlastnú zbierku (collection) so svojimi súbormi na transkripciu. Nahrávanie, import obrazov jednorazovo, je možné do veľkosti 500 MB. Ak je objem importovaných obrazov väčší, obrazy je možné rozdeliť do viacerých súborov a importovať ich postupne. Väčšie súbory obrazov je možné nahrať, importovať aj s použitím FTP klienta, napríklad *WinSCP*, tiež cez URL alebo DFG Viewer

METS. Obrazy sa môžu nahrať ako PDF i JPG, TIFF a i. Zbierka importovaných obrazov, vytvorených skenovaním listov Andreja Kmeťa, má 11.7 GB v rozlíšení 300 dpi.

Naše skúsenosti ukazujú, že pred importom je vhodné skontrolovať digitálne obrazy, ich kvalitu, ostrosť, priesvity z opačnej strany listu, úplnosť, orientáciu strán a pod. Po určitých skúsenostiach sme importovali aj veľké súbory vo formáte PDF cez rýchlejší jednoduchý softvér *WinSCP*.

Segmentácia

Po importe súborov na server sa musí vykonať na serveri automatická segmentácia. Pri segmentácii textu a obrazov musí byť klient pripojený na aplikáciu na serveri. Segmentácia znamená, že sa obraz rukopisného textu dokumentu, ktorý je zatiaľ na serveri ako obraz, rozdelí automaticky na bloky, oblasti, riadky textu. Ak je to potrebné, môžu sa urobiť manuálne korekcie. Ide pritom napríklad o usporiadanie, spájanie a rozdeľovanie blokov, rozširovanie polygónu, úprava základnej linky pod riadkom, ohraničenia segmentu a pod. Segmentácia je pre samotnú transkripciu kľúčová. Kvalitne skenované strany s ostrým rukopisom sa segmentujú spravidla bezchybne. Avšak niekedy je potrebné po segmentácii starostlivo kontrolovať, prípadne upraviť manuálne poradie častí textu (*TR-Text regions*), poradie riadkov (*Lines reading orders*), linky a polygóny vytvorené strojom (umelou inteligenciou).

Tréning stroja HTR

Stroj *Transkribus Expert Client* sa trénuje, cvičí, vlastne učí najprv na stranách, ktoré sú vybraté do cvičného súboru. Stroj opakovane, napr. v 50 cykloch, číta jednotlivé strany cvičného súboru a postupne identifikuje znaky, ktoré nevie jednoznačne určiť, alebo ktoré vznikli chybnou transkripciou strán v súbore *ground truth*.

Transkribus si najprv vytvára model na stranách cvičného súboru. Znaky, ktoré považuje stroj za chybné, zaradí medzi chybné znaky cvičného súboru. To je v štatistike hodnota *CER Train Set*. Stroj HTR musí byť najprv vyškolený pre danú ruku. Spravidla by mal učiaci sa stroj vidieť 100 príkladov každého znaku, ktorý sa nachádza v dokumente, čo je zvyčajne približne na 50 stranách manuálne pripraveného cvičného súboru (Mühlberger et al. [2016]).

Po vycvičení modelu na stránkach, ktoré boli vybraté do cvičného súboru, *Transkribus Expert Client* automaticky použije naučený model vytvorený na stránkach cvičného súboru na jeho overenie na stránkach, vybratých do overovacieho súboru. Overovací súbor, tzv. *Validation set* slúži na praktické vyskúšanie modelu. Ku textu v overovacom súbore stroj pristupuje opakovane zakaždým, akoby to robil prvýkrát a aplikuje pritom model, ktorý sa „naučil“ na cvičnom súbore. Na konci tohto procesu máme k dispozícii model na automatický prepis rukopisu. Pre hodnotenie presnosti transkripcie vytvoreného modelu je najdôležitejšia hodnota, ktorá vyjadruje chybovosť transkripcie znakov vo validačnom, overovacom súbore. To je hodnota *CER Validation Set*.

Z importovanej zbierky sa teda podľa určitého algoritmu vyberie vzorka strán (súbor dát, tzv. dataset), ktorá slúži na učenie stroja a vytvorenie modelu pre určitý typ rukopisu. Na to je potrebné ukázať stroju správne príklady textu. Stroj sa podľa cvičnej, tréningovej sady naučí vzory písma a slov. Ak je zbierka textov od viacerých rúk, je potrebné vybrať primeranú veľkosť cvičnej i testovacej vzorky podľa rúk. Výber strán je možné urobiť podľa určitého algoritmu aj automaticky tak, aby bola vzorka pripravená podľa určitých strán a aby obsahovala asi 20 000 slov. Cvičný, tréningový súbor sa tvorí priamo v expertovom

editore klienta platformy *Transkribus Expert Client* jednak lokálne, jednak aj na serveri. V podstate je potrebné pozorne a veľmi presne prepísať rukopis v editore podľa riadkov, nič neopravovať. Text treba prepisovať podľa súdobého jazykového úzu a gramatiky, aj s chybami a podľa ďalších inštrukcií a návodov, ktoré sú k tejto operácii k dispozícii. Poradie častí textu, označovanie tagmi, výber a redakciu kľúčových slov, deskriptívne metadáta a pod. určuje autor transkripcie a tvorca modelu transkripcie. Výsledok transkripcie je potom viditeľný a zhodnotený na testovacom súbore. Ak je výsledok uspokojivý, možno automaticky transkribovať ďalšie súbory alebo celú zbierku. Jednoducho, po skončení procesu učenia stroja a vytvorení modelu je model k dispozícii vlastníkovi, ktorý ho môže sám používať alebo zdieľať s inými používateľmi a aplikovať na akýkoľvek dokument. Údaje o správnom a nesprávnom čítaní sa stávajú základom modelu.

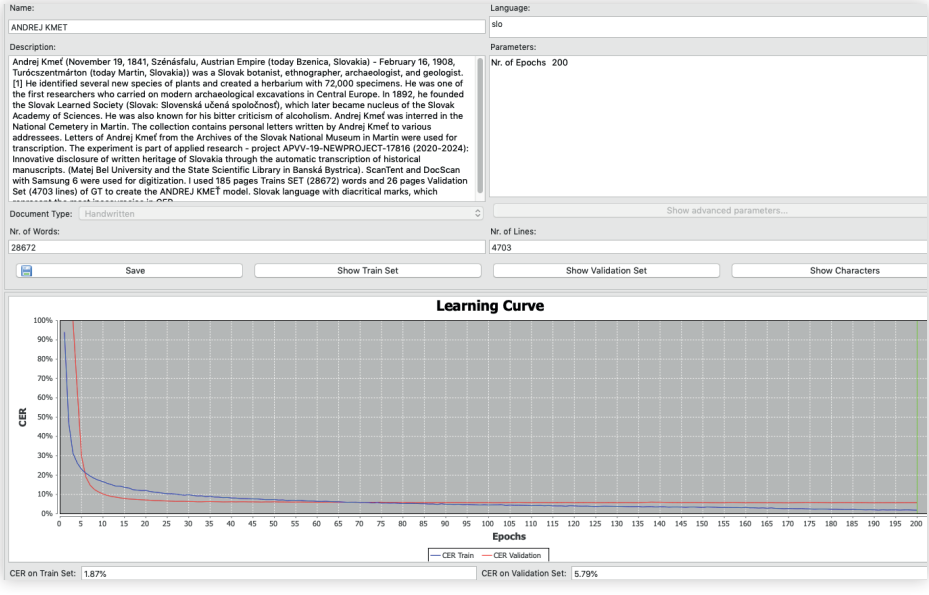
Automatická transkripcia

Automatická transkripcia slúži ako základ pre vedecké editovanie, v ktorom je možné text korigovať, explicitne pridávať ďalšie dáta, kontextové informácie, dešifrovanie dát, určovať tagy, dávať poznámky, metadáta, anotácie, opravy diakritiky, skratky, malé a veľké písmená, paleografické spracovanie, ligatúry a pod.

Automatickú transkripciu sme urobili po spustení tréningu a testovania. Použili sme vlastný model transkripcie a spustili sme transkripciu s použitím HTR+.

Výsledkom učenia v automatickej transkripcii textu rukopisu Andreja Kmeťa bol spočiatku excelentný výsledok CER 1,37 % v tréningovom sete a CER 1,76 % v testovacom sete. Tréningový set obsahoval 29 411 slov a 4 573 riadkov. Model sme použili na ďalšie listy a tie sme opravili tak, aby boli v kvalite *ground truth*.

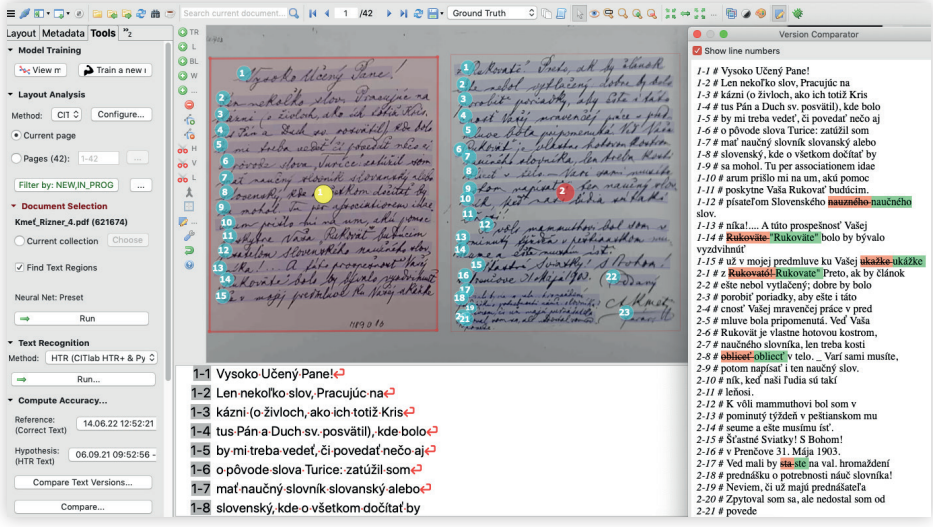
V procese zoznamovania sa s platformou *Transkribus Expert Client* a cez naše pokusy a omyly sme so strojom HTR v roku 2019 prešli od chybovosti 22,81 % v roku 2018 ku chybovosti 1,76 % v roku 2021. Efektívnosť transkripcie sa výrazne zlepšila potom,



Obr. 5 Obrazovka s údajmi po automatickej konverzii s použitím vlastného modelu ID 36009

keď sa stal dostupný stroj HTR+. Spočiatku sme pracovali len s cvičnými súbormi, ktoré neboli v kvalite *ground truth*. Základný cvičný transkribovaný súbor mal 50 strán. Pomerne ľahko sme tento základný súbor zväčšovali až na 185 strán tak, že sme so starším modelom transkribovali ďalšie strany. Tie sme opravovali a pridávali do cvičného súboru. Nové strany sme sa usilovali opraviť čo najpresnejšie do kvality *ground truth*.

Napokon sme vytvorili zo stránok v kvalite *ground truth* spomínaný model č. 36009, ktorým sa dajú dosiahnuť dobré až excelentné výsledky transkripcie, a to v závislosti na kvalite obrazov, ostrosti písma, kvalite rukopisu a kvalite segmentácie. Predbežne môžme konštatovať, že veľká časť chýb transkripcie sa týka interpunkcie. Podrobná analýza príčin nepresností bude predmetom ďalšieho výskumu, rovnako ako výskum korelácie medzi kvalitou skenovania a segmentácie vzhľadom na kvalitu transkripcie.



Obr. 6 Segmentácia textu, transkripcia v editore Transkribus a výsledok automatickej transkripcie

Transkripcia fraktúry (švabachu)[24]

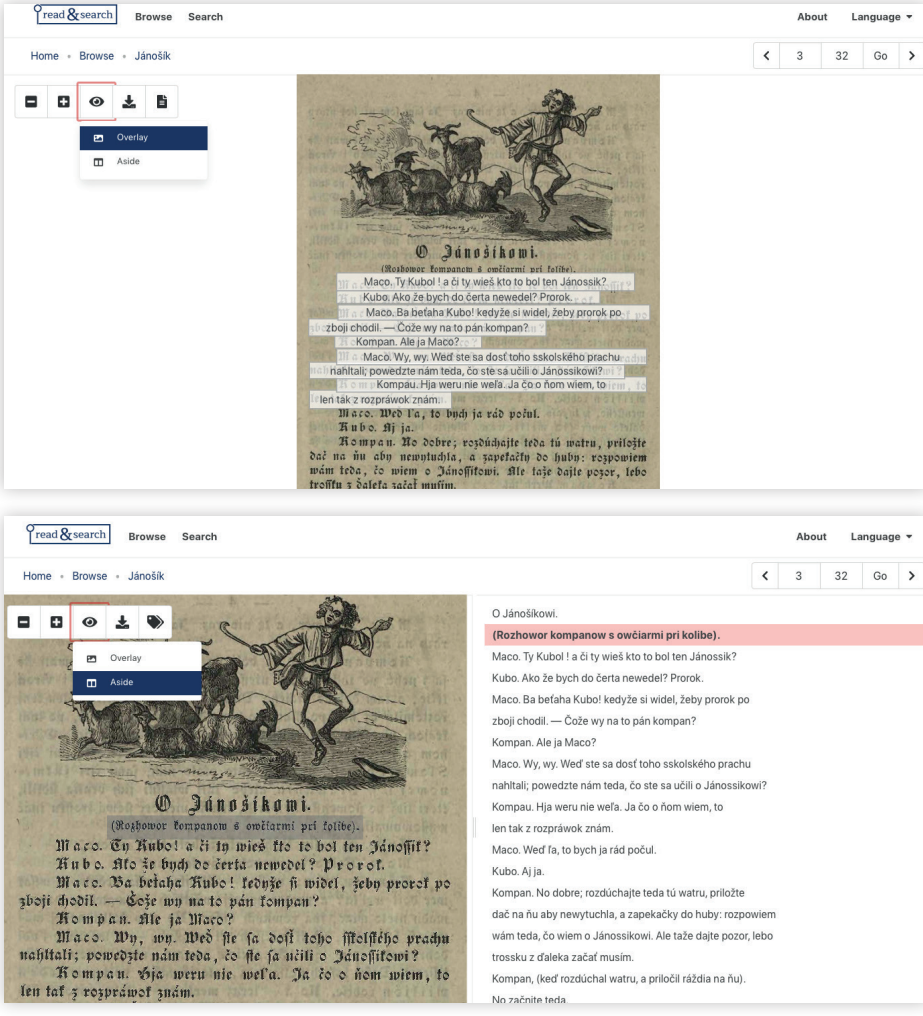
Experiment sa týkal aplikácie umelej inteligencie na automatickú transkripciu slovenskej a českej fraktúry, švabachu (Voit 2006). Fraktúra je typ gotického tlačeného písma[25], ktoré sa vo veľkej miere používalo od 15. storočia aj v českých a slovenských knihách, novinách a časopisoch v novoveku a neskôr, prakticky až do 50. rokov 20. storočia.

V rámci vzdelávania v predmete digitalizácia na Sliezskej univerzite v Ústave bohemistiky a knihovníctva sme uplatnili nástroje umelej inteligencie *Transkribus Expert*

[24] Vynikajúcim znalcom písma je Petr Voit. V jeho prácach sú ukážky variant písma českých historických tlačí, ktoré je rozhodne potrebné preskúmať z hľadiska transkripcie.

[25] *Gotické písmo* malo niekoľko druhov. Napríklad francúzska *textúra* s veľmi ostrým lomom a štíhlou stavbou, talianska širšia a okrúhlejšia *rotunda* s miernejším lomením oblúkov, zmiešané písmo – *bastarda*, v Nemecku *švabach* – písmo širších, oválnejších tvarov a *fraktúra* – písmo užších a špicatejších tvarov s ozdobnými úponkami. Vynálezom kníhtlače (v roku 1450 Johannom Gutenbergom) sa tento druh písma veľmi rozšíril najmä v nemecky hovoriacich krajinách.

Client na prípravu pravdepodobne prvej mimoriadne úspešnej transkripcie slovenského a českého tlačeného textu – fraktúry – historických Moravských novín, Opavského besedníka, slovenskej publikácie Jánošík. Pripravili sme modely transkripcie slovenskej a českej fraktúry (tabuľka 3). V cvičnom súbore sme dosiahli chybovosť CER 0,39 %. Pre praktické využitie tohto modelu je však rozhodujúca vyššia hodnota – 0,44 %, dosiahnutá na validačnom súbore.



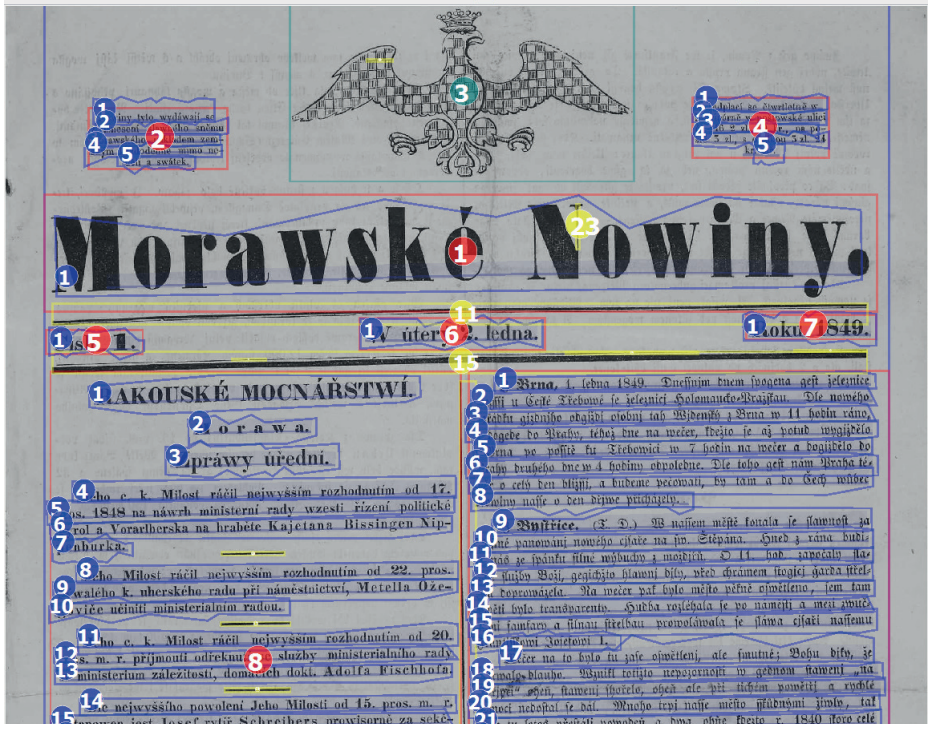
Obr. 7 Transkripcia tlače J. N. Bobulu Jánošík zverejnená v read & search (hore vedľa textu, dolu cez text)

Tab. 3 Transkripcia fraktúry (švabachu)

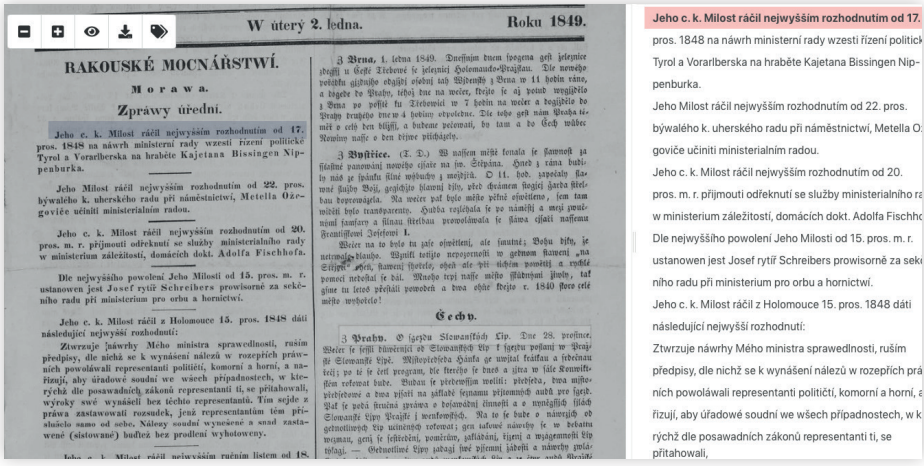
Dátum	Metóda	Tréningový súbor		Validačný súbor		CER		ID modelu
		strany	riadky	strany	riadky	tréningový	validačný	
20210824	OCRbase 29418	7	8092	1	888	0,20 %	0,91 %	36160
20210905	OCRbase 29418	9	11231	4	1179	0,18 %	1,07 %	36353
20210912	OCRbase 29418	17	20805	5	2252	0,39 %	0,44 %	36550
20210913	OCRbase 36550	7	2462	3	276	0,03 %	1,78 %	36607

Fraktúru v slovenských a českých historických tlačiach sme odteraz schopní transkribovať s presnosťou okolo 99 %. V našom prípade je presnosť 99,56 %. Chybovosť je 0,44 %.

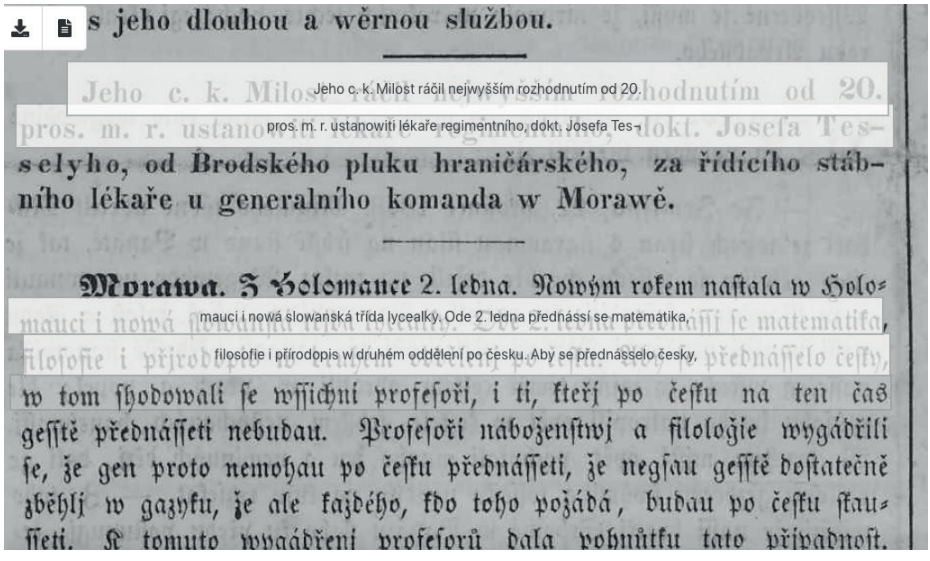
Výsledky prepisu zlomu českého textu sú dostupné po prihlásení sa do platformy *Transkribus* v zbierke FRAKTURA_CZ (114429, Vlastník) a na internete v beta verzii prehliadača *read & search*.



Obr. 8 Ukážka segmentácie Moravských novín 1849 (antikva a fraktúra)



Obr. 9 Ukážka transkripcie Moravských novín 1849 s použitím vlastného modelu transkripcie



Obr. 10 Výrez transkripcie a zobrazenie textu cez transkribovaný text v read & search

Ďalší výskum

V ďalšom výskume bude vhodné zamerať pozornosť na tieto oblasti: a) výber a štandardný popis rozsiahlejších slovacikálnych rukopisných zbierok európskeho a národného významu, b) digitalizácia vybratých historických dokumentov podľa plánu experimentov s cieľom potvrdiť alebo zlepšiť doteraz známe postupy a hodnoty vzhľadom na nasledujúci proces segmentácie textu a automatickú transkripciu (korelácia medzi rôznymi podmienkami a kvalitou skenovania a transkripciou, c) dôkladná analýza a popis výsledkov segmentácie textov, d) zdieľanie digitálnych dokumentov s archívmi a inými inštitúciami, ktoré ich budú môcť používať podľa vlastnej úvahy ako náhradu papierových dokumentov, e) tvorba modelov, tréning a analýza modelov automatickej transkripcie podľa novovekých a moderných zbierok a jazykov (najmä slovenčina, čeština, maďarčina, latinčina, nemčina, poľština), f) overenie a zhodnotenie použiteľnosti hotových, dostupných modelov transkripcie z výskumu v projekte READ, g) zoznámenie sa s najlepšou praxou automatického rozpoznávania textov historických dokumentov v Európe, najmä v Nemecku, Rakúsku, Španielsku, Maďarsku, Veľkej Británii, Fínsku, Holandsku, Srbsku, využitie informácií a skúseností na Slovensku, h) automatická transkripcia podstatnej časti rukopisnej Laučekovej[26] zbierky a jej virtualizácia, teda virtuálna jedna digitálna prezentácia zväzkov, ktoré sa nachádzajú na geograficky rozličných miestach (Slovenská národná knižnica v Martine, Slovenský národný archív v Bratislave, Univerzitná knižnica v Bratislave, Országos Széchenyi Konyvtár v Budapešti), i) výskum možností zvýšenia efektívnosti rozpoznávania rukopisných textov a textov historických dokumentov prostredníctvom platformy *Transkribus* a súvisiacich nástrojov, j) sprístupnenie transkribovaných a interpretovaných zbierok cez digitálny repozitár širokej verejnosti, k) tvorba dokumentácie, ktorá bude slúžiť pre archívy, knižnice, akademické pracoviská ako aj fyzické osoby na automatickú transkripciu textov, l) vybudovanie *kabinetu digital humanities* so zameraním na transkripciu historických dokumentov.

Záver. Efektívnosť platformy Transkribus

Naše skúsenosti overené experimentami potvrdzujú, že rukopisy je možné automaticky transkribovať, pričom chybovosť môže byť veľmi nízka a výsledok je excelentný. Výsledky transkripcie sú čitateľné a možno ich exportovať v rôznych formátoch – DOC, TXT, PDF, TEI, METS, ďalej editovať, redigovať, korigovať a použiť.

V experimente sme pri *rukopise* Andreja Kmeťa dosiahli presnosť 94,21 % pri chybovosti znakov (CER) 5,79 %. V transkripcii *tlačenej fraktúry* sme dosiahli presnosť 99,56 % pri chybovosti znakov 0,44 %.

Z hľadiska vnímania, porozumenia a použitia transkribovaného textu vo všeobecnosti podľa autorov platformy *Transkribus* platí, že: a) ak sa striktne počíta chybovosť „slov“ a ak chybovosť slov je do 30 %, tak text je pre človeka ešte pochopiteľný a použiteľný,

[26] Martin Lauček (* 12. máj 1732, Martin – † 9. február 1802, Skalica) bol slovenský evanjelický kňaz, prekladateľ a náboženský spisovateľ. Je autorom monumentálneho rukopisného diela Collectanea. Ide asi o 24 zväzkov a približne 20 000 strán. Svojim obsahom sú Collectanea neoceniteľným zdrojom poznatkov a informácií k dejinám evanjelickej cirkvi a prameňom k histórii protestantizmu. Našim cieľom je jednak zhromaždiť všetky dostupné zväzky a vytvoriť jednu virtuálnu verejne dostupnú digitálnu zbierku. Ďalej analyzovať texty a pokúsiť sa o ich automatickú transkripciu a zverejnenie pre všetkých.

b) ak sa striktne počíta chybovosť „znakov“, a ak chybovosť znakov je do 15 %, tak text je ešte pre človeka pochopiteľný a použiteľný.

Platforma *Transkribus* je skvelou pomôckou pre svedomitých a trpezlivých bádateľov, ktorých v žiadnom prípade nenahradí, ale podstatne uľahčí doladenie transkripcie cez editovanie a korektúry výsledkov. Platforma nie je, a sotva niekedy bude, určená len pre *„klikavcov“,* teda používateľov, ktorí sú zvyknutí viac „klikať“ ako trpezlivo inovovať.

Poďakovanie

PhDr. Márii Halmovej, *Mgr. Viere Varínskej* a *PhDr. Anne Peťovej* za pomoc pri snímaní rukopisov Andreja Kmeťa v etnografickom múzeu v Martine.

Oľge Kuchtovej z Banskej Štiavnice za pomoc pri zisťovaní informácií o živote a podmienkach pôsobenia Andreja Kmeťa v Prenčove.

Mgr. Márii Bôbovej, PhD. zo Štátnej vedeckej knižnice v Banskej Bystrici za pomoc a spoluprácu pri manuálnej transkripcii a segmentácii strán pre cvičný model a transkripciu listov Andreja Kmeťa.

Lucii Valjentovej, študentke knihovníctva zo 4. ročníka Ústavu bohemistiky a knihovníctva Sliezskej univerzity v Opave za pomoc pri transkripcii českej fraktúry.

Alešovi Drahotušskému za poskytnutie novín z Digitálnej knižnice Štátnej vedeckej knižnice v Ostrave.

Zoznam bibliografických odkazov

KATUŠČÁK, D., I. NAGY, M. BÔBOVÁ, P. KUNC, A. KURHAJCOVÁ, P. MALINIAK, M. MIKUŠKOVÁ, L. NIŽNÍKOVÁ, I. POLÁKOVÁ, B. SNOPKOVÁ a O. TOMEČEK. (2019) SKRIPTOR Projekt APVV-19-NEWPROJECT-17816 (2020–2024). Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov. [Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts]. Organizácie: Univerzita Mateja Bela v Banskej Bystrici (zodpovedný riešiteľ doc. Imrich Nagy, PhD.) a Štátna vedecká knižnica v Banskej Bystrici – partner (garant prof. PhDr. Dušan Katuščák, PhD.).

ADAM MATTHEW DIGITAL, 2018. *Handwritten text recognition: artificial intelligence transforms discoverability of handwritten manuscripts*, [cit. 2. 10. 2021]. Dostupné z: *www.amdigital.co.uk/products/handwritten-text-recognition*.

BÔBOVÁ, M., 2021. Projekt Skriptor, keď stroj sa stáva žiakom. In: *Vedecká online konferencia NON SCHOLAE, SED VITAE DISCIMUS*, dňa 7. júna 2021 v gescii ŠVK v Prešove.

DROBAC, S., 2020. OCR and post-correction of historical newspapers and journals (Doctoral dissertation). Helsinki: University of Helsinki, 2020. ISBN 978-951-51-6511-4 (paperback), ISBN 978-951-51-6512-1 (PDF), [cit. 10. 6. 2022]. Dostupné z: *https://helda.helsinki.fi/bitstream/handle/10138/319496/OCRandpo.pdf?sequence=1 & isAllowed=y*.

HODEL T., D. SCHOCH, C. SCHNEIDER a J. PURCELL, 2021. General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. *Journal of Open Humanities Data*, *7*, 13. [cit. 1. 10. 2022]. Dostupné z: *https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.46/.*

HOLLÝ, K., 2013. *Veda a slovenské národné hnutie: snahy o organizovanie a inštitucionalizovanie vedy v slovenskom národnom hnutí v dokumentoch 1863–1898.* Bratislava: Historický ústav SAV v Typoset Print, s. r. o., 2013.

HOLLÝ, K., 2015. *Andrej Kmeť a slovenské národné hnutie: Sondy do života a kreovanie historickej pamäti do roku 1914.* Bratislava: Veda, Historický ústav SAV, 2015. 279 s. ISBN 978-80-224-1480-7.

HOTSON, H. a T. WALLNIG (eds.), 2019. Reassembling the Republic of Letters in the Digital Age. Göttingen: Göttingen University Press, 2019. 470 s. [COST Action IS1310; 2014–2018. ISBN 978-3-86395-403-1. DOI: *https://doi.org/10.17875/gup2019-1146.* [cit. 1. 10. 2022] Dostupné z: *https://www.univerlag.uni-goettingen.de/handle/3/isbn-978-3-86395-403-1*.

KATRENIAK, M. (2022). Automatická transkripcia rukopisných historických textov na príklade vybraných kanonických vizitácií. Dostupné z:*https://opac.crzp.sk/?fn=detailBiblioForm & sid=BDC-2D20A28F62792149F199B8B08.*

KATUŠČÁK, D., 2008. Súčasný stav formovania stratégie digitalizácie na Slovensku. In: *Kolokvium knihovních a informačních pracovníků zemí V4+.* 6.–8. července 2008, Brno, ČR. Elektronický sborník, s. 30–46.

KATUŠČÁK, D., 2021. Pochybná hodnota za veľa peňazí? *Kultúrny kyslík*. 2021, č. 2, s. 14–17. [cit. 3. 10. 2021]. ISSN 1339-6919. Dostupné z: *https://via-cultura.sk/kulturny-kyslik-2-2021/.*

KATUŠČÁK, D. a M. KATUŠČÁK, 2011. Základná koncepcia národného projektu digitálna knižnica. *Knižnica*, 2011, **12**(2), 6–10. [cit. 2. 10. 2021] Dostupné z: *https://www.snk.sk/images/snk/casopis_kniznica/2011/februar/06.pdf.*

KATUŠČÁK, D., 2011a. *Digitálna knižnica a digitálny archív*. Národný projekt. Operačný program informatizácie spoločnosti OPIS2. Implementácia 2010–2015. Martin: Slovenská národná knižnica, 2011. [Kompletný projekt k žiadosti o nenávratný finančný príspevok zo štrukturálnych fondov Európskej únie ca 4000 s.].

KATUŠČÁK, D., 2011b. Národný projekt digitálna knižnica a digitálny archív. *Bulletin Slovenskej asociácie knižníc*. Bratislava: SAK, 2011. 38 s. [Opis projektu] Dostupné z: *http://dusan.katuscak.net/2011/12/02/digitalna-kniznica-a-digitalny-archiv-opis2/.*

KATUŠČÁK, D., 2011c. Situační zpráva o národním projektu SNK Digitální knihovna a digitální archiv. In: 12. konference Archivy, knihovny, muzea v digitálním světě 2011. Praha: SKIP, 30. listopadu a 1. prosince 2011 v konferenčním sále Národního archivu v Praze, Archivní 4, Praha 4 – Chodovec. [cit. 2. 10. 2021] Dostupné z: *http://old.skipcr.cz/dokumenty/akm-2011/Katuscak.pdf.*

KATUŠČÁK, D., 2021. Progress in making available blackletters typefaces and handwritten written heritage using artificial intelligence. Preprint. *Researchgate*. 2021, 25 s.

KOVÁČOVÁ, K., 2022. [bakalárska práca] Výběr pozoruhodných rukopisných sbírek Jesenicka. [cit. 2. 10. 2022]. Dostupné z: *https://is.slu.cz/th/bum3h/FPF_BP_2022_53474_Kovacova_Klara.pdf.pdf.*

KIŠŠ, M., 2018. Rozpoznávání historických textů pomocí hlubokých neuronových sítí. Brno, 2018. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Michal Hradiš, Ph.D.

MARTÍNEK, J., L. LENC a P. KRÁL, 2020. Building an efficient OCR system for historical documents with little training data. Neural Computing & Applications 32, 17209–17227 (2020). [cit. 2. 10. 2021] Dostupné z: *https://doi.org/10.1007/s00521-020-04910-x.*

MINISTERSTVO KULTÚRY SLOVENSKEJ REPUBLIKY, 2019. Revízia výdavkov na kultúru. Priebežná správa. Október 2019. Kap. 4.4 Projekt digitalizácie, s. 75–78. [cit. 2. 10. 2021] Dostupné z: *https://www.culture.gov.sk/wp-content/uploads/2019/12/Revizia_vydavkov_na_kulturu_priebezna_sprava_compressed.pdf*.

MINISTERSTVO KULTÚRY SLOVENSKEJ REPUBLIKY, 2020. Revízia výdavkov na kultúru. Záverečná správa. Júl 2020. Kap. 4.9 Digitalizácia kultúrneho dedičstva, 132–139. [cit. 2. 10. 2021] Dostupné z: *https://www.culture.gov.sk/wp-content/uploads/2020/10/Revizia_vydavkov_na_kulturu_-_zaverecna_sprava_compressed.pdf*.

MÜHLBERGER, G., 2016. READ (Recognition and Enrichment of Archival Documents) – 2016–2019. [Projektová štúdia]. [cit 6. 10. 2021.] Dostupné z: *https://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019.*

MÜHLBERGER, G., L. SEAWARD, M. TERRAS, S. ARES OLIVEIRA, V. BOSCH, M. BRYAN, S. COLUTTO, 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, E.-M. LANG, S. LAUBE, G. LEIFERT, G. LOULOUDIS, R. McNICHOLL, J.-L. MEUNIER, J. MICHAEL, E. MÜHLBAUER, N. PHILIPP, I. PRATIKAKIS, J. PUIGCERVER PÉREZ, H. PUTZ, G. RETSINAS, V. ROMERO, R. SABLATNIG, J.-A. SÁNCHEZ, P. SCHOFIELD, G. SFIKAS, C. SIEBER, N. STAMATOPOULOS, T. STRAUSS, T. TERBUL, A. H. TOSELLI, B. ULREICH, M. VILLEGAS, E. VIDAL, J. WALCHER, M. WEIDEMANN, H. WURSTER a K. ZAGORIS, 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, **75**(5), 954–976. Dostupné z: *https://doi.org/10.1108/JD-07-2018-0114*.

MÜHLBERGER, G., J. ZELGER a D. SAGMEISTER, 2014. User-driven correction of OCR errors: combining crowdsourcing and information retrieval technology. In: ANATONACOPOULOS, A. & K. U. SCHULZ. (Eds.), *DATeCH'14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* Madrid, Spain, 19–20 May 2014 (s. 53–56). New York, NY: Association for Computing Machinery. Dostupné z*: https://doi.org/10.1145/2595188.2595212*.

MÜHLBERGER, G., S. COLUTTO a P. KAHLE [2016, Preprint]. Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars. The Model of a Transcription & Recognition Platform (TRP). Dostupné z: *https://www.academia.edu/8601748/Preprint_Handwritten_Text_Recognition_HTR_of_Historical_Documents_as_a_Shared_Task_for_Archivists_Computer_Scientists_and_Humanities_Scholars_The_Model_of_a_Transcription_and_Recognition_Platform_TRP_?bulkDownload=thisPaper-topRelated-sameAuthor-citingThis-citedByThis-secondOrderCitations & from=cover_page*.

MÜHLBERGER, G., 2002. Digitising instead of mailing or shipping: a new approach to interlibrary loan through customer-related digitisation of monographs. *Interlending & Document Supply*, **30**(2), 66–72. Dostupné z: *https://doi.org/10.1108/02641610210430523*.

NAGY, I., 2021. Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov = The Possibilities of application the method of digital transcription of historical manuscript texts in the process of accessing the archival fonds. *Slovenská archivistika*. Bratislava: Ministerstvo vnútra Slovenskej republiky, 2021, **51**(2), 53–67. ISSN 0231-6722. Dostupné z: *https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/SA%202-2021,%20roc.%2051.pdf*.

POOLE, A. H., 2017. The Conceptual Ecology of Digital Humanities. *Journal of Documentation*, 2017. **73**(1), 91–122. [cit. 3. 10. 2021]. Dostupné z: *https://www.academia.edu/27862789/The_Conceptual_Ecology_of_Digital_Humanities*.

STROBEL, P. B., S. CLEMATIDE a M. VOLK, 2020. How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3551–3559. Marseille, 11–16 May 2020 c European Language Resources Association (ELRA).

ŠTUDENTSKÁ grantová súťaž SGS/5/2022 (SGS SU Opava). *Tvorba modelu automatické transkripce historického rukopisu s využitím umělé inteligence.* Řešitel: prof. PhDr. Dušan Katruščák, PhD., Ing. I. Kyselová, PhD., od októbra 2022 aj K. Kováčová.

KOVÁČOVÁ, K. a I. KYSELOVÁ, 2022. Robot čte rukopisnou kuchařskou knihu z roku 1667? In: *Študentská vedecká konferencia*. Slezská univerzita v Opavě, 5. apríla 2022.

TOMEČEK, O., 2021. Metales Banskej Bystrice z roku 1820. Reambulácia juhozápadného úseku mestských hraníc spoločných so susedným teritóriom rodiny Radvanských = Metales of the town Banská Bystrica from 1820. Perambulation of the southwest part of town borderline common with neighbouring domain of Radvanský family / Oto Tomeček. *Acta historica Neosoliensia: vedecký časopis pre historické vedy*. Banská Bystrica: Vydavateľstvo Univerzity Mateja Bela – Belianum, 2021, **24**(2), 112–133. ISSN 1336-9148. Dostupné z: *https://www.ahn.umb.sk/tomus-24-num-2-tomecek-o-metales-banskej-bystrice-z-roku-1820-reambulacia-juhozapadneho-useku-mestskych-hranic-spolocnych-so susednym-teritoriom-rodiny-radvanskych/*.

VOIT, P., 2006. *Encyklopedie knihy: starší knihtisk a příbuzné obory mezi polovinou 15. a počátkem 19. století*. Praha 2006. Švabach – Encyklopedie knihy. [cit. 2. 10. 2022]. Dostupné z: *https://www.encyklopedieknihy.cz/index.php/%C5%A0vabach*.

KATUŠČÁK, Dušan. Umelá inteligencia pomáha sprístupňovať písomné dedičstvo. *Knihovna: knihovnická revue*. 2022, **33**(2), 50–77. ISSN 1801-3252.