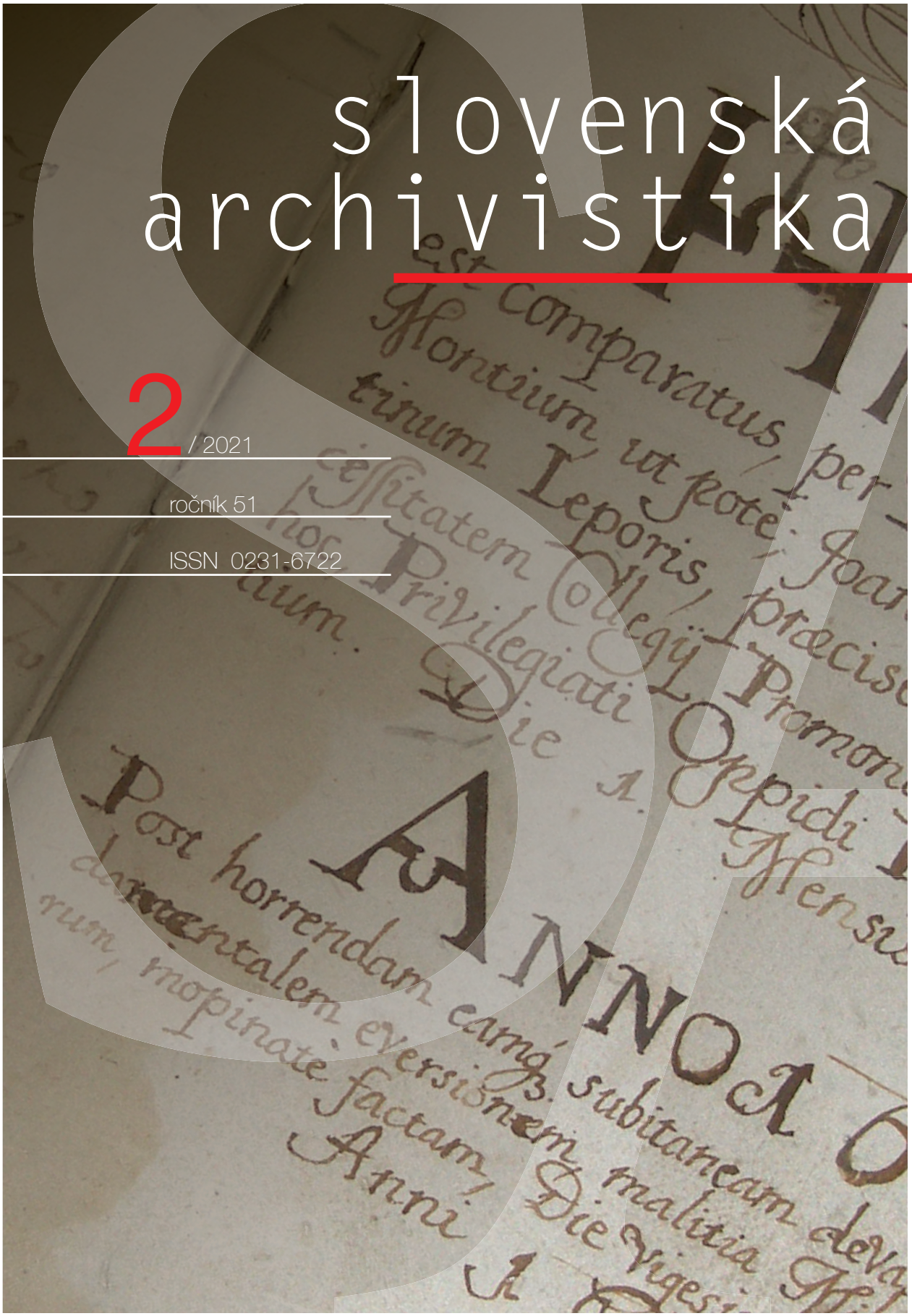
slovenská archivistika

2 / 2021

ročník 51

ISSN 0231-6722



SLOVENSKÁ ARCHIVISTIKA

ROČNÍK 51
2021
Číslo 2

ODBOR ARCHÍVOV A REGISTRATÚR
MINISTERSTVA VNÚTRA SLOVENSKEJ REPUBLIKY
A
MINISTERSTVO VNÚTRA SLOVENSKEJ REPUBLIKY
SLOVENSKÝ NÁRODNÝ ARCHÍV

Vydavateľ/Publisher
Ministerstvo vnútra Slovenskej republiky
odbor archívov a registratúr a Slovenský národný archív

OBSAH

ZOZNAM AUTOROV ŠTÚDIÍ A ČLÁNKOV

Ing. Dušana Greschová, Slovenský národný archív, Drotárska cesta 42, P. O.BOX 115, 840 05 Bratislava 45, e-mail: dusana.greschova@minv.sk

Mgr. Juraj Hirčák, Mincovňa Kremnica, š. p., Štefánikovo nám. 25/24, 967 01 Kremnica, e-mail: archiv@mint.sk

Mgr. Ľuboš Janaček, Štátny archív v Bratislave, Križkova 7, 811 04 Bratislava, e-mail: lubos.janacek@minv.sk

Karol Kantek, Komenského 23, 900 01 Modra, e-mail: kantek@zoznam.sk

PhDr. Eva Kowalská, DrSc., Historický ústav Slovenskej akadémie vied, P. O. Box 198, Klemensova 19, 814 99 Bratislava, e-mail: eva.kowalska@savba.sk

doc. Mgr. Imrich Nagy, PhD., Katedra histórie Filozofickej fakulty Univerzity Mateja Bela, Tajovského 40, 974 01 Banská Bystrica, e-mail: imrich.nagy@umb.sk

MOŽNOSTI APLIKÁCIE METÓDY DIGITÁLNEJ TRANSKRIPCIE HISTORICKÝCH RUKOPISNÝCH TEXTOV PRI SPRÍSTUPŇOVANÍ ARCHÍVNYCH FONDOV*

IMRICH NAGY

Nagy, I.: The Possibilities of Application the Method of Digital Transcription of Historical Manuscript Texts in the Process of Accessing the Archival Fonds. Slovenská archivistika, Vol. 51, 2021, No 2, p. 53-67.

The article discuss the method of automatic transcription of the historical manuscript texts and the possibilities of its use in the process of accessing the archival fonds. Application of this procedure is examined on a sample of a contemporary archival aid – the catalog of correspondence of the Koháry family from the fond in the State archives in Banská Bystrica, compiled by J. Csákos in the years 1944 – 1945. It depicts the individual phases of preparation of the model for its automatic transcription. Finally, it analyses the success and effectiveness of automatic transcription. At the end, the author states very good results which open the possibility for transcription of a complete numerical catalog of Koháry's correspondence, on the basis of which it will be possible to realize its further research, or even its edition.

Key words: correspondence catalog, Transkribus, automatic trascription, archival fond

Zdôvodňovať význam digitalizácie pri ochrane a sprístupňovaní písomného dedičstva (resp. akejkoľvek písomnej informácie) už v dnešnom svete, ktorý sa pri konfrontácii s dopadmi globálnej pandémie rýchlo adaptoval na masové využitie digitálnych komunikačných kanálov, azda nikomu netreba. Napriek tomu, že v rámci operačných programov Európskej únie bola nakúpená špičková technika a vybudované národné digitalizačné centrá, pri digitalizácii archívnych fondov sú citeľné veľké rezervy a hovoriť o funkčnom elektronickom archíve je nutné stále ešte v budúcom čase. Jednou z príčin sú aj nespracované fondy bez opisov, inventárov, katalógov a registrov, čo sťažuje orientáciu a vyhľadávanie v nich. V niektorých prípadoch sú síce k dispozícii dobové, historické archívne pomôcky a katalógy, tieto však nezodpovedajú súčasným štandardom a vyhľadávanie v nich je neraz skôr prácnou bádateľskou činnosťou. Práve tu však moderné technológie využívajúce umelú inteligenciu môžu podať veľmi účinnú pomocnú ruku.

V našej štúdii predstavíme metódu Handwritten Text Recognition (HTR+), t. j. metódu automatickej transkripcie historických rukopisných textov, ktorú vyvinulo konzorcium pod vedením Güntera Mühlbergera z Univerzity v Innsbrucku v rámci

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

projektu Horizon 2020 Recognition and Enrichment of Archival Documents (READ).[1] Jej základný kameň tvorí platforma Transkribus, ktorá využíva obrovský potenciál najnovších technológií tzv. neural engine, t. j. strojového učenia. Transkribus za krátky čas dokázal pritiahnuť v odborných inštitúciách mnohých štátov Európy značnú pozornosť.[2] V súčasnosti sa ďalej rozvíja a ponúka na komerčné, vedecké a vzdelávacie využitie v rámci združenia READ-COOP SCE, ktoré má v súčasnosti už 86 členov z 24 krajín, pričom medzi nimi prevládajú popri univerzitách práve archívy a knižnice.[3] Jediným zástupcom z krajín V4 je Univerzita Mateja Bela v Banskej Bystrici, na ktorej v spolupráci so Štátnou vedeckou knižnicou v Banskej Bystrici v súčasnosti v rámci projektu aplikovaného výskumu APVV „Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov" overujeme možnosti aplikácie nástroja Transkribus na rôzne typy historických rukopisných textov z územia Slovenska, vyvíjame nové modely na ich automatickú transkripciu a overujeme efektívnosť, funkčnosť, ako aj praktickú využiteľnosť metódy HTR+ v prostredí slovenských pamäťových inštitúcií, s ktorými sme nadviazali spoluprácu (Literárny archív Slovenskej národnej knižnice v Martine, Diecézny archív Banskobystrickej diecézy a odbor archívov a registratúr sekcie verejnej správy Ministerstva vnútra SR).

Vďaka spolupráci s odborom archívov a registratúr sme z prostredia štátnych archívov získali indikatívne návrhy a zoznamy archívnych dokumentov, ktoré jednotlivé archívne pracoviská odporučili na spracovanie metódou HTR+. Pri výbere dokumentu pre pilotnú analýzu a aplikáciu nástroja Transkribus padla voľba práve na dobovú archívnu pomôcku – číselný katalóg korešpondencie z archívneho fondu Koháry-Coburg, ktorý spracoval bratislavský archivár János József Csákós v rokoch 1944 – 1945.[4] Katalóg obsahuje regesty 6 632 listov rôzneho rozsahu (v niektorých prípadoch obsahujúce aj výťahy, resp. preklady originálnych listov in extenso) v maďarskom jazyku spísané v rukopise v podobe tabuliek na 4 140 stranách formátu A3 (250 × 400 mm) v zošitovej väzbe.[5] Hoci samotný rukopis je teda veľmi mladého dáta (typologicky ide o moderné kurzívne písmo), zaujímavý a hodnotný je sprístupnením mimoriadne cenného a rozsiahleho fondu korešpondencie rodu Koháry-Coburg, ktorý pre svoju nespracovanosť ostáva takmer nevyužívaný. Orientácia v samotnom katalógu, ktorý je radený podľa pôvodného číslovania listov v rodovom archíve, a to bez ďalšej logic-

[1] https://cordis.europa.eu/project/id/674943 [cit. 2021-08-23].

[2] Transkribus získal v roku 2020 aj ocenenie EÚ Horizon Impact Award. [cit. 2021-08-23] Dostupné na internete: https://cordis.europa.eu/article/id/422311-horizon-impact-award-2020-awards-eu-funded-projects-with-the-greatest-societal-impact. Na Slovensku ako prvý upozornil na pozoruhodné výsledky projektu aj s analýzou možností ich aplikácie a využitia v našich podmienkach prof. Katuščák. Pozri: KATUŠČÁK, Dušan. Digital humanities a automatická transkripcia rukopisných textov. In. *ITlib: Informačné Technológie a Knižnice*, 2020, roč. 24, č. 1, s. 6-16. ISSN 1335-793X.

[3] https://readcoop.eu/members/ [cit. 2021-08-23].

[4] OTRUBA, Štefan. Štátny archív v *Banskej Bystrici Sprievodca po archívnych fondoch II.* Bratislava : Slovenská archívna správa, 1969, s. 17.

[5] Štátny archív v Banskej Bystrici (ďalej len ŠA BB), fond Koháry-Coburg (1241) 1321 – 1945, časť IV., číselný katalóg korešpondencie.

kej súvislosti, je mimoriadne ťažkopádna a bez využitia ďalších nadväzujúcich dobových archívnych pomôcok de facto nemožná. Neprehľadnosť katalógu môže odrádzať bádateľov, hoci samotný fond korešpondencie obsahuje obzvlášť cenné informácie najmä k vojenským dejinám 17. a 18. storočia, osobitne k protiosmanským vojnám, ale aj detailný pohľad do zákulisia každodenného života vysokej šľachty a cisárskeho dvora v uvedenom období, či hospodárskeho a ekonomického fungovania panstva Kohárovcov.

Ciele experimentu

Pri voľbe uvedeného katalógu na pilotnú aplikáciu nástroja Transkribus sme vychádzali z nasledujúcich hypotéz:

1. veľký rozsah uniformného rukopisu zaručí mimoriadnu efektivitu využiteľnosti vytrénovaného modelu automatickej transkripcie;
2. možnosť voľby rozsiahleho súboru vstupných dát pre trénovanie modelu automatickej transkripcie zabezpečí jeho vyladenie, ktoré sa prejaví v nízkej miere chybovosti a praktickej bezproblémovej čitateľnosti, resp. zrozumiteľnosti digitálneho výstupu;
3. digitálne textové výstupy z automatickej transkripcie umožnia nielen fulltextové vyhľadávania, ale aj ďalšie štruktúrovanie informácií obsiahnutých v rukopise;
4. výstupy z transkripcie môžu byť dostupné a využiteľné na vedecké, výskumné a vzdelávacie účely.

V nasledujúcom texte ponúkneme popis a vysvetlenie krokov pri jednotlivých fázach práce s nástrojom Transkribus, ako aj prvé čiastkové analýzy a hodnotenia vo vzťahu k vyššie uvedeným hypotézam.

Digitalizácia dokumentu

Transkribus pracuje s digitálnymi grafickými súbormi vo formátoch: JPEG, PNG a TIFF, resp. univerzálnym súborovým formátom PDF obsahujúcim grafické súbory. Tu sa vraciame späť k téme digitalizácie a potreby mimoriadne výkonnej (a drahej) digitalizačnej techniky, ktorá sa odôvodňuje vysokými nárokmi na kvalitu (rozlíšenie) výslednej digitálnej kópie. V prípade Transkribusu však táto premisa celkom neplatí: jeho vývojármi bol otestovaný a odladený pre spracovanie fotografických záznamov dokumentov vyhotovených mobilnými telefónmi, ktoré už bežne disponujú dostatočne kvalitnou optikou.[6] Na tento účel bol vyvinutý jednoduchý statív s neutrálnym osvetlením v podobe tzv. ScanTentu.[7]

---

[6] K tomu bližšie pozri: KLEBER, Florian et. al. *Mass Digitization of Archival Documents using Mobile Phones.* HIP2017: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing November 2017, s. 65-70. DOI: 10.1145/3151509.3151526 [cit. 2021-08-24] Dostupné na internete: https://resolver.obvsg.at/urn:nbn:at:at-ubtuw:3-3361.

[7] Popis a princíp fungovania ScanTentu je dostupný na webstránke združenia Read-Coop. [cit. 2021-08-24] Dostupné na internete: https://readcoop.eu/scantent/.

OBRÁZOK 1 – *Práca so ScanTentom.*
Foto: M. Nagyová

S využitím tejto pomôcky sme pomocou mobilného telefónu iPhone 11 Pro zdigitalizovali kompletný rozsah vyššie špecifikovaného dokumentu. Vyhotovili sme celkom 2 275 snímok s rozmermi 4032 × 3024 pixelov pri rozlíšení 192 DPI zachytávajúcich dvojstránku originálneho dokumentu. Pri prvej, resp. poslednej strane jednotlivých zošitov originálu obsahuje snímka len jednu stranu umiestnenú na kontrastnom (čiernom) povrchu pracovnej plochy ScanTentu, čo viedlo k badateľnému preexponovaniu dotknutých snímok. Súčasťou našej pracovnej hypotézy sa teda stalo aj overenie použiteľnosti snímok vyhotovených na konkrétnom zariadení s uvedenými charakteristikami pre Transkribus.

Tvorba modelu pre automatickú transkripciu

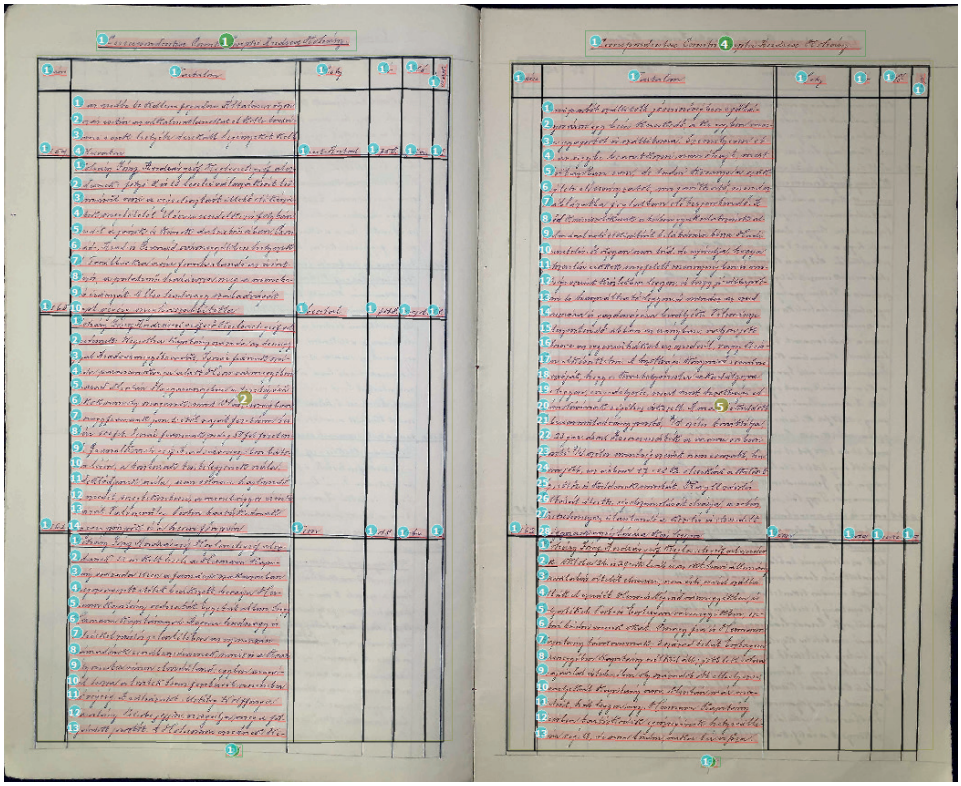
Ako sme uviedli, platforma Transkribus je založená na mechanizme tzv. neural engine, t. j. strojovom učení. V našom prípade, interpretované zjednodušene, sa obrazovej predlohe znaku na origináli priraďuje konkrétny alfanumerický znak. Transkribus tento prevod porovná s Ground Truth, t. j. skutočnou hodnotou, ktorú by mal tejto predlohe priradiť. Ak zistí nesprávnu hodnotu, vyradí ju z množiny možných riešení. Opakovaním tohto procesu, teda učením, sa množina možných riešení, ktoré Transkribus priraďuje obrazovej predlohe znaku postupne zmenšuje až na hodnotu rovnú skutočnej hodnote znaku. Inými slovami „stroj sme naučili, ako má čítať konkrétny obrazový znak“. Tomu potom už pri jeho každom ďalšom výskyte bude priraďovať rovnakú hodnotu. Ako z tejto interpretácie vidno, pre aplikáciu Transkribusu na konkrétny dokument je v zásade irelevantné, aký je starý, v akom jazyku je napísaný, akým rukopisom či akým typom písma. Pre úspešný výsledok automatickej transkripcie je potrebné stroj „naučiť čítať“ konkrétny rukopis na základe pripravenej vzorky obsahujúcej už presne priradené alfanumerické znaky. Inými slovami: musíme si pripraviť bezchybný prepis vzorky textu, na základe ktorého vytrénujeme model na automatickú transkripciu. Autori Transkribusu odporúčajú pre takúto vzorku rukopisu rozsah okolo 15 000 slov.[8] V našom prípade išlo o 29 snímok, čiže digitálnych obrazov obsahujúcich prvých 53 strán zo 4 140 stranového rukopisu.

Príprava Ground Truth vzorky na platforme Transkribusu pozostáva po importe snímok na vzdialený server do vyhradenej a chránenej osobnej zbierky používateľa[9]

[8] MUEHLBERGER, Guenter et al. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In *Journal of Documentation*, 2019, roč. 75, č. 5, s. 959. ISSN 0022-0418.

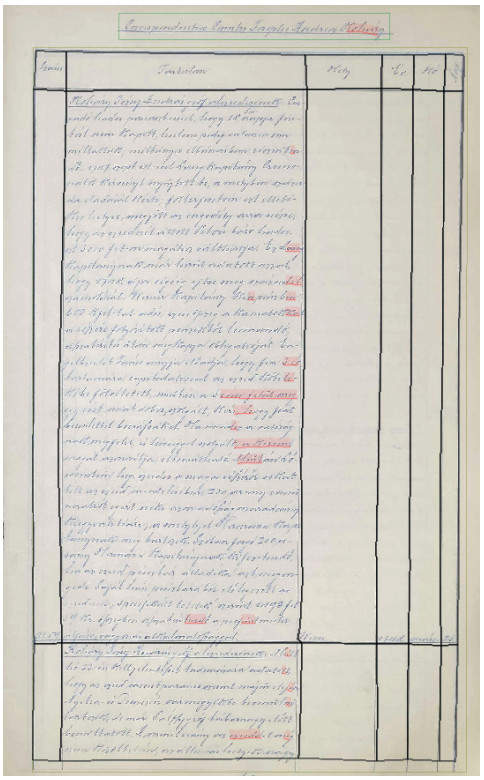
[9] Ide vlastne o variáciu cloudového riešenia, do ktorého sa registrovaný používateľ prihlasuje cez svoje užívateľské meno a heslo. Nespornou výhodou je, že sa k svojej zbierke môže prihlasovať

cez rozhranie aplikácie z dvoch fáz. Prvou je tzv. segmentácia textu, t. j. rozlíšenie štruktúry a orientácie textu, jeho vymedzenie do blokov a riadkov a určenie poradia čítania zistených blokov a v nich jednotlivých riadkov. Tento proces je možné automatizovať a na používateľovi potom zostáva len kontrola a oprava (napr. spresnenie hranice riadku, zmena poradia čítania a pod.). V prípade našej vzorky však bola situácia komplikovanejšia, keďže originál rukopisu mal formu tabuľky. V tomto prípade bolo potrebné rozdeliť segmentáciu na dva kroky: v prvom sme manuálne vymedzili bloky textu, ktoré netvoria súčasť tabuľky (záhlavie, poznámky na margu, paginácia) a vymedzili blok textu pre tabuľku, ktorý sme horizontálnym a vertikálnym delením rozčlenili na jednotlivé stĺpce a riadky. Upraviť bolo potrebné aj poradie čítania jednotlivých buniek (Transkribus postupuje automaticky zhora nadol a zľava doprava). Jednotlivé textové rámce (riadky/stĺpce/bunky tabuľky) je možné označiť metadátami. My sme takto označili záhlavie a podľa druhu informácií stĺpce (číslo listu, obsah, lokalita, datovanie).



OBRÁZOK 2 – *Vyznačené textové rámce a hranice riadkov s určením poradia čítania na dvojstránke Csákósovho rukopisného katalógu.*
Zdroj: Transkribus.

z akéhokoľvek miesta z počítača s nainštalovanou aplikáciou a s prístupom na internet, pričom je možný aj paralelný prístup viacerých používateľov k tej istej zbierke súčasne.



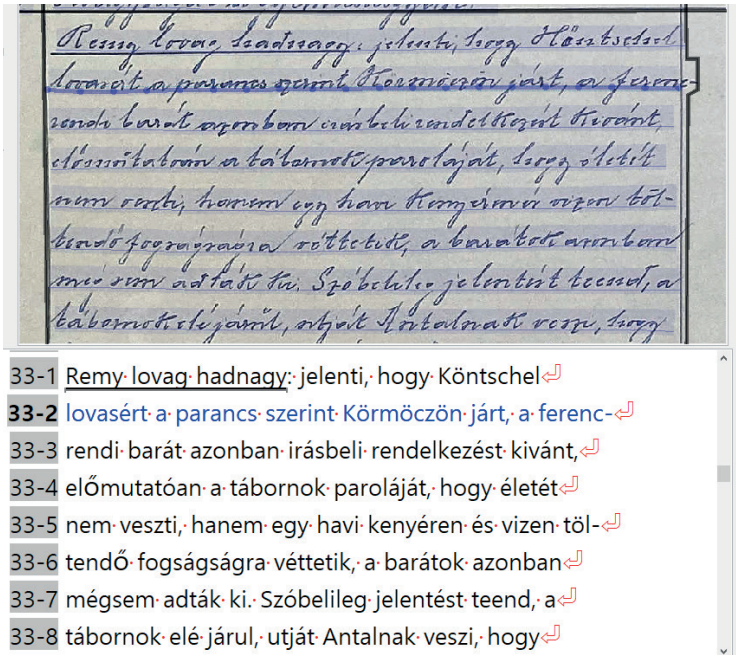
OBRÁZOK 3 – *Strana s chybne vyznačenými hranicami riadkov po automatickej segmentácii.*
Zdroj: Transkribus.

Takto zadefinované bloky textu je možné, za predpokladu, že tabuľka má unifikovanú podobu, skopírovať na všetky snímky v danej zbierke. V našom prípade mali unifikovanú podobu (veľkosť) iba stĺpce, tie sme preniesli aj na ostatné snímky. Dokončenie segmentácie – vyznačenie hraníc riadkov – vykonal v preddefinovaných rámcoch tabuľky už automaticky Transkribus. Manuálne bolo potrebné ich iba korigovať. V tejto fáze vyskočil problém so spomínanými preexponovanými snímkami, na ktorých Transkribus pri automatickej segmentácii nedokázal vyznačiť takmer žiadne hranice riadkov. Časová náročnosť úplnej segmentácie jednej dvojstrany bola priemerne 10 minút.

Príprava Ground Truth vzorky sa zavŕšila presným prepisom originálu do platformy Transkribusu. V tejto fáze je potrebné každému vyznačenému riadku priradiť alfanumerické znaky exaktne zodpovedajúce originálnemu rukopisu, ktoré kopírujú všetky jeho omyly a nepresnosti, pričom akékoľvek odchýlenie od originálu sa môže neskôr prejaviť na chybovosti modelu.

Vzorka Ground Truth sa nakoniec rozdelí odporúčane v pomere 10 : 1 na cvičný súbor (Training set) a overovací súbor (Validation set). Trénovanie modelu (a jeho následné overenie) Transkribus opakuje – pre efektívny model je štandardne nastavených 50 cyklov (epochs).[10] Na cvičnom súbore sa Transkribus „učí“, t. j. číta pri každom cykle rovnaké strany, ale chybné čítania znakov sa pri každom nasledujúcom cykle vyradia z množiny možných riešení. Inými slovami „pamätá si, kde sa pomýlil“. Tieto údaje o správnom a nesprávnom čítaní sa stávajú základom modelu. Po vytrénovaní modelu na stránkach, ktoré boli vybraté do cvičného súboru, ho Transkribus automaticky použije na stránkach cvičného súboru. Overovací súbor, tzv. validation set, slúži na praktické odskúšanie modelu. K textu v overovacom súbore pristupuje zakaždým, akoby to robil prvýkrát a aplikuje pritom to, čo sa „naučil“ na cvičnom súbore. Na konci tohto procesu máme k dispozícii model pre automatický prepis rukopisu J. Csákósa aj s jeho základnými charakteristikami (pozri obr. 5).

[10] Počet cyklov si používateľ môže nastaviť aj na inú hodnotu 100, 200, 1000…, čo môže mať vplyv na spresnenie modelu.



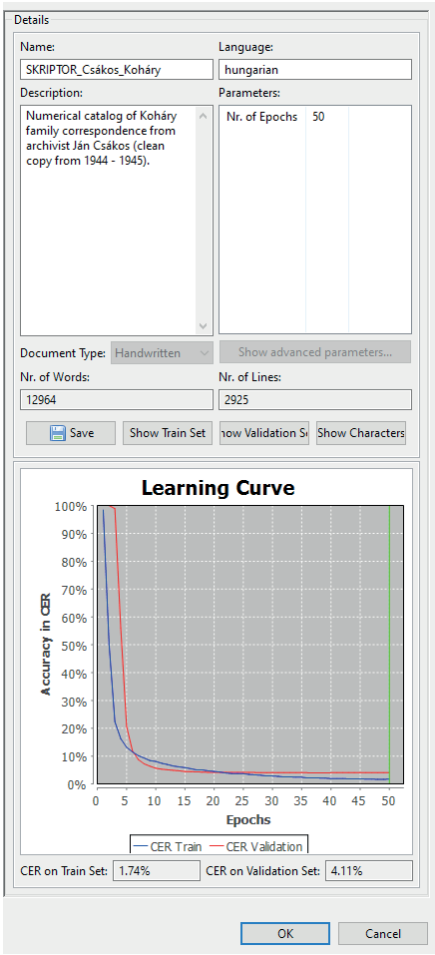
OBRÁZOK 4 – *Ukážka prepisu rukopisného originálu v Transkribuse. Zdroj:* Transkribus.

Analýza a hodnotenie modelu pre automatickú transkripciu

V charakteristike modelu máme k dispozícii aj údaje o percente chybovosti, ktoré Transkribus automaticky vypočítaval pri každom cykle trénovania modelu, pričom pod touto chybovosťou sa rozumie percento nesprávne určených alfanumerických znakov (CER, t. j. character error rate) z celého textu. Tento štatistický ukazovateľ počítal osobitne pri trénovaní – na vyhodnotenie chýb, ktorých sa dopustil pri čítaní cvičného súboru (CER on Train Set) a pri čítaní overovacieho súboru (CER on Validation Set).

Na obr. 5 vidíme v grafickej podobe vývoj chybovosti pri čítaní cvičného súboru (modrá krivka) a osobitne overovacieho súboru (červená krivka), číselne je uvedený výsledný údaj po skončení trénovania modelu, ktorý je v našom prípade na úrovni 1,74 % pri cvičnom súbore a 4,11 % pri overovacom súbore. Zaujímavejší je údaj pri overovacom súbore, ktorý ukazuje schopnosť Transkribusu zvládnuť „prečítanie“, resp. automatický prepis ľubovoľnej strany skúmaného rukopisu. Konkrétne náš údaj hovorí o tom, že 95,81 % znakov z cvičného súboru bolo určených v rámci procesu automatickej transkripcie bezchybne. Ako hraničná chybovosť, dokedy je možné hovoriť o zmysluplnosti automatickej transkripcie, sa uvádza 10 % CER. Náš výsledok pod 5 % CER sa pri rukopisoch považuje za vynikajúci výsledok.[11] Lepšie hodnoty

[11] MUEHLBERGER, ref. 8, s. 962.



OBRÁZOK 5 – *Charakteristika modelu pre automatickú transkripciu Csákósovho katalógu.*
Zdroj: Transkribus.

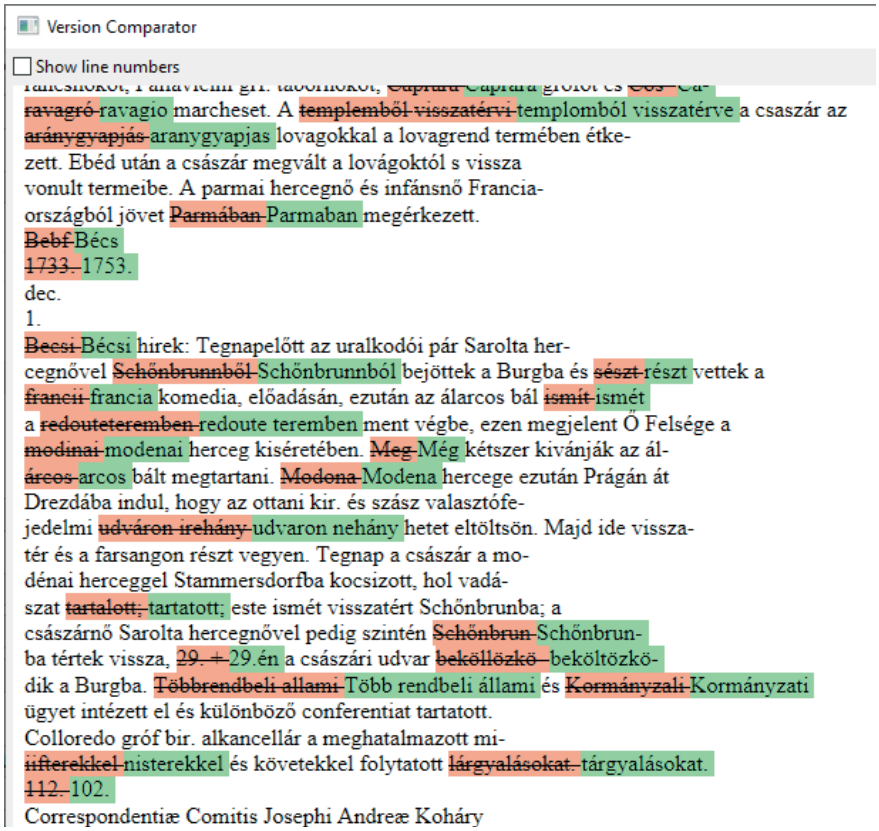
okolo 1 – 2 % CER sa zvyknú dosahovať len pri modeloch pre historické tlačoviny.[12]

Model sme, samozrejme, chceli overiť na ďalšej dávke 28 snímok, ktoré obsahovali strany 54 – 105 Csákósovho rukopisu. Aj pred automatickou transkripciou týchto strán je potrebné vykonať segmentáciu textu. Rovnakým postupom, aký sme popísali vyššie pri trénovaní modelu, sme segmentovali všetkých 28 strán až na úroveň určenia hraníc riadkov a kontroly ich poradia. Potom sme spustili proces automatickej transkripcie založený na nami vytrénovanom modeli. Text, ktorý sme dostali, bol relatívne čitateľný a zrozumiteľný, obsahoval však chyby najmä v interpunkcii, diakritike, čísliciach, ale aj pri niektorých písmenách (badateľná bola najmä tendencia automatického prepisu „ll“ namiesto správneho „k“ a pod.).

Rozhodli sme sa teda overiť chybovosť na úrovni konkrétnych strán. K tomu sme ich potrebovali korigovať na verziu Ground Truth. Zjavná bola časová úspora (približne v pomere 1 : 10) v porovnaní s úplným manuálnym prepisom do verzie Ground Truth pri tvorbe modelu pri prvých 53 stranách rukopisu. Na záver sme nechali Transkribus porovnať našu korigovanú verziu prepisu (Ground Truth) s prvotným výstupom automatickej transkripcie založenom na našom modeli. Výsledky vybraných porovnávaných strán prinášame v tabuľke 1.

Z výsledkov vyplýva, že CER sa pohybuje v rozmedzí od 1,67 % (excelentný výsledok) po 8,12 % (použiteľný výsledok). Priemer CER zo všetkých porovnávaných strán 5,26 % s malou odchýlkou (1,15 percentuálneho bodu) zodpovedá CER nášho modelu 4,11 %. Vo výsledkoch porovnania sa objavuje aj údaj WER (word error rate) udávajúci percento chybovosti slov, ktorý sa štandardne používa pri hodnotení úspešnosti metódy OCR (optical character recognition), ktorá sa používa v komerčných softvéroch na rozpoznanie tlačeného textu z obrázkových súborov (napr. Abbyy FineReader). Vidíme, že WER sa pri analyzovaných stranách

[12] STRÖBEL, Phillip – CLEMATIDE, Simon. (2019). Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images. Utrecht: Digital Humanities 2019. Posted at the Zurich Open Repository and Archive, University of Zurich. [cit. 2021-08-26]. Dostupné na internete: https://doi.org/10.5167/uzh-177164



OBRÁZOK 6 – *Ukážka porovnania automaticky transkribovaného textu s jeho korigovanou verziou.*
Zdroj: Transkribus.

pohybuje v porovnaní s CER vo vyšších hladinách od 7,95 % po 24,94 %, čo pri druhom údaji fakticky znamená, že každé štvrté slovo na príslušnom snímku je chybne prepísané. Takáto chybovosť by už spochybňovala účelnosť automatickej transkripcie. Musíme si však uvedomiť, že najviac chybných slov obsahuje väčšinou chyby späté s interpunkciou (chýbajúca alebo naopak nadbytočná bodka, čiarka, dvojbodka a pod.), resp. s diakritikou (krátka samohláska namiesto dlhej, resp. naopak),[13] ktoré nemajú takmer žiadny vplyv na zrozumiteľnosť textu. Pri metóde HTR+, ktorá sa využíva na platforme Transkribus, sa teda preferuje sledovanie CER pred WER.[14]

Hľadali sme odpoveď aj na otázku, prečo zaznamenávame taký veľký rozptyl miery chybovosti na jednotlivých stranách. Rukopis Csákósa je relatívne unifikovaný, hoci bezo sporu vykazuje aj nejaké vybočenia, čo je však pre tvorbu úspešného mo-

[13] Porov. obr. 6 s vyznačenými chybnými slovami.

[14] HODEL, Tobias et al. General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. In. *Journal of Open Humanities Data*, 2021, roč. 7, č. 13, s. 4. [cit. 2021-08-25]. Dostupné na internete: DOI: https://doi.org/10.5334/johd.46.

Strany	WER	CER
Page 1	7,95	1,67
Page 2	12,6	4,03
Page 3	21,41	6,31
Page 4	12,41	2,86
Page 7	15,29	3,47
Page 8	10,8	2,38
Page 11	19,12	4,66
Page 12	20,14	4,83
Page 13	23,66	6,53
Page 14	23,02	4,45
Page 15	22,29	8,12
Page 16	22,25	6,08
Page 18	20,63	6,38
Page 19	17,96	4,45
Page 20	17,03	4,42
Page 21	23	7,77
Page 22	19,71	6,34
Page 24	23,02	6,08
Page 27	24,94	6,49
Page 28	24,86	7,91
PRIEMER	**19,105**	**5,262**

TABUĽKA 1 – *Prehľad Word error rate (WER) a character error rate (CER) na vybraných stranách automaticky transkribovaného textu.*

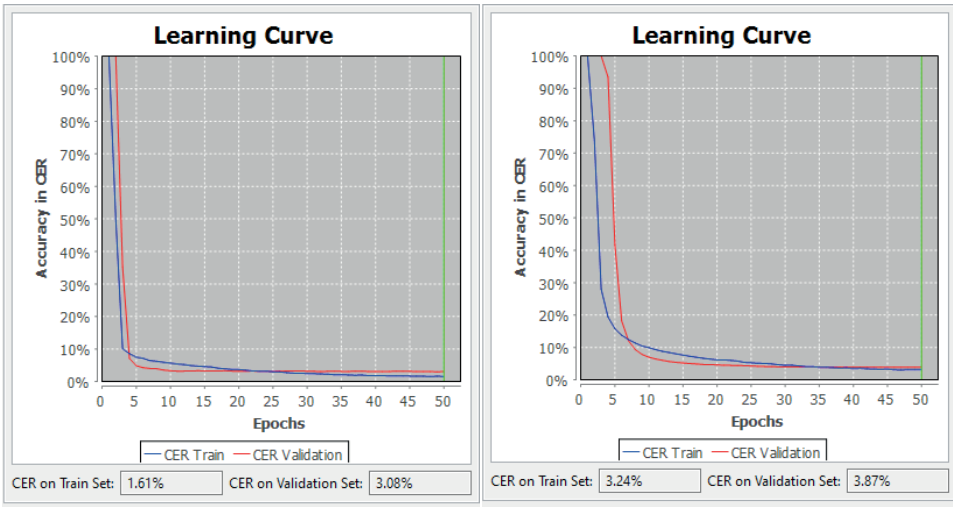
delu na automatickú transkripciu skôr pozitívne.[15] Výraznejší dopad na úspešnosť transkripcie mala samotná výraznosť (sýtosť) písma a najmä kvalita jeho digitálnej snímky. Ako sme upozornili, medzi snímkami boli aj výrazne preexponované zábery. Ak si však teraz budeme chcieť overiť našu čiastkovú pracovnú hypotézu, že si Transkribus s týmito stranami nedokáže poradiť, zistíme, že strany z inkriminovaných snímok patria skôr k stranám transkribovaným s nízkou mierou chybovosti (CER v rozpätí od 2,38 % do 4,45 %). Je to paradoxné, keď si uvedomíme, že Transkribus pri automatickej segmentácii týchto strán de facto zlyhal.[16]

Výsledkom prípravy podkladov pre komparáciu úspešnosti automatickej transkripcie bol aj ďalší súbor Ground Truth, ktorý sa núkal pre zdokonalenie už existujúceho modelu. Predpokladali sme, že pridaním ďalších strán do cvičného súboru, dosiahneme lepší výsledok v podobe hodnôt CER na cvičnom aj overovacom súbore. Experiment sme urobili dvomi spôsobmi: pri prvom sme trénovali model (Model A) s druhým súborom Ground Truth obsahujúcim 28 snímok s využitím nášho prvého modelu ako základného (base model). Pri druhom spôsobe sme urobili trénovanie úplne nového modelu (Model B) s využitím oboch súborov Ground Truth, ktoré spolu obsahovali 57 snímok.

V prvom prípade sme dosiahli vytrénovanie nového modelu s CER 1,61 % na cvičnom súbore a 3,08 % na overovacom súbore. Ak si tieto údaje porovnáme s chybovosťou prvého modelu, vidíme, že pri oboch súboroch došlo k zlepšeniu o 0,13 percentuálneho bodu (cvičný súbor), resp. 1,03 percentuálneho bodu (overovací súbor). Ukazovatele CER pri modeli vytvorenom druhým spôsobom vykázali naopak vyššiu mieru chybovosti (3,24 % pri cvičnom súbore a 3,87 % pri overovacom súbore). Pri zdokonaľovaní modelu je teda vhodnejším riešením využívať už existujúci model ako základný. Zároveň tým potvrdzujeme platnosť hypotézy č. 2: metódu automatickej transkripcie historických rukopisov možno odporučiť v prvom rade pre dokumenty od identického pôvodcu, ktorých rozsah výrazne presahuje odporúčaný minimálny počet 15 000 slov pre tvorbu modelu. Pri dobrých výsledkoch základného modelu (CER ≤ 3 %) však už nemožno očakávať jeho vylepšovaním (rozširovaním vzorky Ground Truth pre trénovanie modelu) výrazné zlepšenie.

[15] HODEL, ref. 14, s. 2.

[16] Môže to byť podnet na ďalší výskum korelácie kvality skenovania a efektívnosti tvorby modelu.



OBRÁZOK 7 – *Porovnanie rozšírených modelov pre automatickú transkripciu Csákósovho katalógu. Vľavo model A, vpravo model B.*
Zdroj: Transkribus

TABUĽKA 2 – *Ukážka výstupu automatickej transkripcie Csákósovho katalógu korešpondencie Kohárvovcov vo formáte docx.*

Correspondentiæ Comitis Josephi Andreæ Kohárv

Szám	Tartalom	Hely	Év	Hó	Nap.
2159.	it az ezredbe be kellene fogadni. Altalános ujon czozás esetén az alkalmatlanokat el kelle bocsá-tani s ezek helyébe derekább legényeket kell fölavatni	Szent-Antal	1750.	jun.	8.
2160	Koháry József András gróf Keglevich gróf alez-redesnek: folyó 8. és 10. levelei válaszaként tu-domásul veszi a végrehajtást illető előkeszü-letek megtételét. Udvari rendelkezés folytán ezredét a jászok és kúnok districtusában, Cson-grád-, Arad-és Csanád vármegyékben helyezik el. Továbbiakra nézve fenntartandó az érint-kezés: a palatinus határozza meg a menete-lés irányát. Allio hadnagy szabadságát szept. végeig meghosszabbította.	Ebental	1748.	szept.	18.

Na základe zistených údajov predpokladáme, že pri automatickej transkripcii ďalších strán Csákósovho katalógu založenej už na novom modeli, dosiahneme lepšie

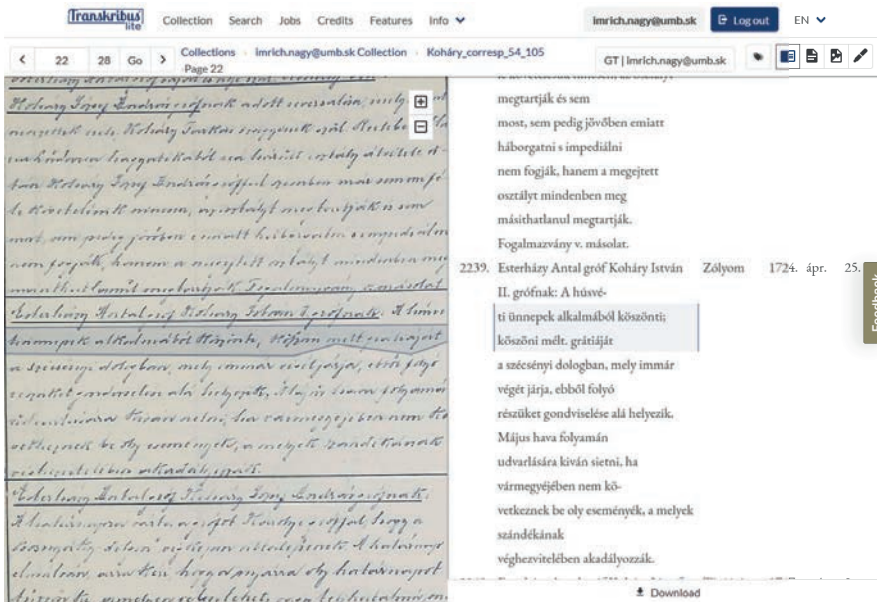
výsledky, t. j. bezprostredne použiteľný digitálny text. Ten možno z Transkribusu exportovať v rôznych formátoch (pdf, docx, txt, xlsx) so všetkými výhodami ďalšieho spracovania textu v príslušnom formáte (napr. vyhľadávanie konkrétneho reťazca znakov či celých slov a výrazov).

Potvrdzuje sa nám teda hypotéza č. 3: vo výstupe z automatickej transkripcie je možné fulltextovo vyhľadávať. Transkribus taktiež umožňuje vkladanie metadát, čo sa dá pohodlne urobiť napr. pri korektúre automatického prepisu. Tieto metadáta je možné preniesť do výstupu v docx, kde je potom s príslušnou funkcionalitou programu MS Word jednoduché vygenerovať k textovému súboru index. Ak teda na archivára/bádateľa, ktorý napríklad chce nájsť konkrétny údaj z korešpondencie Kohárovcov, doposiaľ čakala takpovediac sizyfovská niekoľkodňová úloha prechádzať všetky dobové pomôcky od začiatku do konca a prácne si vypisovať jeho výskyt, po realizácii automatickej transkripcie môže mať výsledok k dispozícii takmer okamžite. Formát pre MS Excel zas umožňuje štruktúrovanie dát, resp. ich zoraďovanie napr. podľa roku, miesta, ale aj akejkoľvek ďalšej doplňujúcej informácie (napr. pôvodcu, adresáta atď.). S tým súvisí de facto aj potvrdenie hypotézy č. 1: efektivitu využiteľnosti vytrénovaného modelu možno jednoducho vyjadriť pomerom času venovaného príprave Ground Truth pre trénovanie modelu voči času ušetrenému pri vyhľadávaní najrozličnejších informácií z automaticky transkribovaného celého dokumentu. Opravy chýb v transkribovanom texte možno prirovnať k redakčnej práci pri príprave edície prameňov.

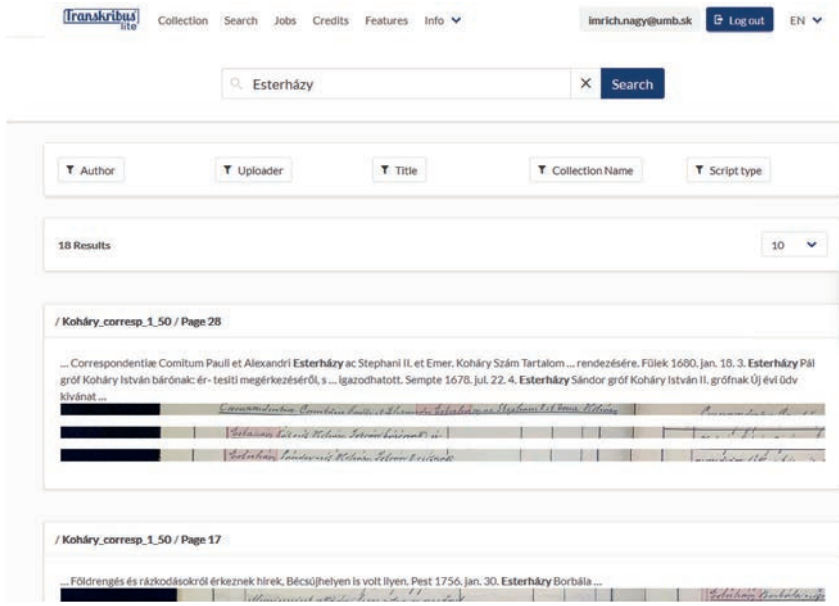
Sprístupňovanie výsledkov automatickej transkripcie

Zmyslom digitalizácie písomného dedičstva je nielen jeho ochrana a uchovanie pre budúce generácie, ale aj sprístupnenie pre vedecké, výskumné a vzdelávacie účely čo najjednoduchším informačným kanálom. Pri ochrane a uchovaní budeme využívať možnosti digitálneho repozitára Univerzity Mateja Bela a Štátnej vedeckej knižnice v Banskej Bystrici. Pre sprístupnenie ponúka vhodný nástroj samotný Transkribus. Ide o webové rozhranie „read&search“,[17] ktorý ponúka prezeranie digitalizovaného originálu paralelne s jeho prepisom s graficky zvýraznenou konkordanciou na úrovni riadkov. Súčasťou je aj vyhľadávací nástroj, ktorý príslušné slovo nájde a jeho výskyt vyznačí opäť paralelne (v digitalizovanom origináli a zároveň aj v jeho prepise) v celej zbierke, resp. túto funkcionalitu ponúka aj odľahčená – webová verzia Transkribusu, tzv. Transkribus Lite, do ktorého má prístup vlastník, resp. používateľ zbierky po prihlásení sa rovnakými prihlasovacími údajmi, akými sa prihlasuje do aplikácie Transkribus na počítači. Tu, po nastavení príslušného filtra, je možné súčasne vyhľadávať aj vo viacerých zbierkach, ku ktorým má prihlásený užívateľ prístup. Objektom vyhľadávania sú iba samostatné slová, ktoré možno upraviť zástupnými znakmi pre rozšírenie vyhľadávania aj na alternatívne tvary slova. Napríklad pri zadaní výrazu „Esterh?zy“ sa zobrazí výskyt mena s diakritikou (Esterházy) aj bez nej (Esterhazy). Podobne pri zadaní výrazu „Antal*“ sa zobrazí výskyt slova Antal a zároveň aj výskyt všetkých tvarov slova s príponami (napr. Antalra, Antalban a pod.).

[17] Pozri popis, jednotlivé funkcionality a ďalšie charakteristiky webového rozhrania na stránke združenia READ-COOP. [cit. 2021-08-29] Dostupné na internete: https://readcoop.eu/read search/.



OBRÁZOK 8 – *Prezeranie digitalizovaného originálu a jeho prepisu cez webové rozhranie Transkribus Lite.*
Zdroj: https://transkribus.eu/lite.



OBRÁZOK 9 – *Zobrazenie výsledku vyhľadávania v digitalizovanom originále a súčasne v prepise cez webové rozhranie Transkribus Lite.*
Zdroj: https://transkribus.eu/lite.

Na základe uvedeného môžeme konštatovať aj potvrdenie našej poslednej pracovnej hypotézy č. 4: Transkribus ponúka nástroje, ktoré umožňujú jednoduché, graficky atraktívne a pritom vysoko účelné sprístupnenie výstupov digitalizovaného originálu i výsledkov vyhľadávania v ňom.

Záver

Výsledky prvotnej, experimentálnej fázy aplikácie metódy HTR+ na vybraný dokument – dobovú archívnu pomôcku hodnotíme ako veľmi dobré, oprávňujúce v nej pokračovať. V rámci riešenia projektu preto plánujeme transkribovať Csákósov katalóg v celom rozsahu, a to nielen z časti IV., ale aj časti V. predmetného fondu, kde sa nachádza číselný katalóg ďalšej časti korešpondencie Kohárovcov. J. Csákós ho spracoval v rokoch 1936 – 1937.[18] Obsahuje 19 758 položiek na 600 stranách vo forme voľných dvojhárkov s rozmermi 240 × 340 mm.[19] Pre úplnosť vytvoríme aj model pre rukopis Eugena Neuenschwandera, ktorý okolo roku 1910 zostavil číselný katalóg vo dvoch zväzkoch v celkovom rozsahu 201 strán s rozmermi 235 × 385 mm k položkám č. 1 – 2000 z korešpondencie Kohárovcov nachádzajúcej sa v časti IV. predmetného fondu.[20] Následne bude možné v celej korešpondencii identifikovať početnosť a rozsah listov identických pôvodcov, čo otvára možnosť pre vytvorenie osobitných modelov na automatickú transkripciu pôvodných listov[21] z korešpondencie Kohárovcov s potenciálnym výhľadom ich edície.

[18] OTRUBA, ref. 4, s. 17.

[19] ŠA BB, fond Koháry-Coburg (1241) 1321 – 1945, časť V., číselný katalóg korešpondencie.

[20] ŠA BB, fond Koháry-Coburg (1241) 1321 – 1945, časť IV., číselný katalóg korešpondencie.

[21] V rámci riešenia projektu APVV-19-0456 SKRIPTOR sú naplánované úlohy na overenie aplikácie metódy HTR+ aj v prípade starších rukopisov (napísaných humanistickou kurzívou) s väčšou variabilitou rukopisu. Predbežné experimenty na rukopisoch postily Izáka Abrahamidesa Hrochotského z rokov 1600 – 1601, či kanonických vizitácií Banskobystrickej diecézy z 18. – 19. storočia prinášajú v tomto ohľade sľubné výsledky, o ktorých budú riešitelia projektu v krátkom čase informovať v ďalších výstupoch.

THE POSSIBILITIES OF APPLICATION THE METHOD OF DIGITAL TRANSCRIPTION OF HISTORICAL MANUSCRIPT TEXTS IN THE PROCESS OF ACCESSING THE ARCHIVAL FONDS

NAGY, Imrich

Modern technologies that use elements of the artificial intelligence based on neural engine, offer new possibilities in accessing the historical manuscript texts. So it is the Transkribus platform, developed within the European project Horizon 2020 READ. To verify its functionality, the authors chose a contemporary archival aid – a catalog of correspondence of Koháry family, processed by J. Csákós from the fonds of the State Archives in Banská Bystrica. The numerical catalog offers abstracts and partly transcripts to 6632 letters on 4140 sheets of A3 paper format. The catalog was digitized by scanning with a camera of an iPhone 11 Pro and ScanTent with a resolution of 192 DPI, sufficient for the HTR+ method using Transkribus platform. On a sample of 29 images containing the first 53 pages of the Csákós'catalog, the basic model for the automatic trascription was practised. The success of the model is determined statistically by the CER indicator which marks the ratio of erroneous characters in the automatically generated transcript. The model achieved CER 4,11 % on the verification file pages. In general, a model with CER ≤ 10 % is rated as functional and with CER ≤ 5 % as succesful. Based on the authors'model, they performed an automatic transcription of another 28 images containing the next 50 pages of Csákós'catalog. The average CER value for automatically transcribed pages was 5,26 %. As a part of the experiment, they expanded the sample file for training the model with corrected pages from the automatic transcription. The corrected model achieved CER 3,08 %. Such a model is fully functional for the automatic transcription. On its base it is possible to transcribe the complete catalog of Koháry's correspondence. Transcription outputs will allow further research of correspondence: identification of persons, locations, events, dating and other data.

Translated by Jana Kafúnová

SLOVENSKÁ ARCHIVISTIKA

Ročník 51 – Číslo 2 / 2021

Vydavateľ/Publisher
Ministerstvo vnútra Slovenskej republiky
odbor archívov a registratúr a Slovenský národný archív

Adresa redakcie a administrácia/Address
MV SR Slovenský národný archív,
Drotárska cesta 42, P. O. Box 115, 840 05 Bratislava 45
slovenska.archivistika@minv.sk

Periodicita/Periodicity
Dvakrát ročne/Biannual
(redakčná uzávierka 28. februára a 31. augusta)

Náklad/Print run
150 ks

EV 140/08
ISSN 02316722

Návrh obálky/Cover design
Nora Nosterská – Perecká kniha z rokov 1681 – 1781, Štátny archív v Trenčíne,
pracovisko Archív Bojnice

Tlač/Print
Centrum polygrafických služieb, p. o. Ministerstva vnútra SR