Aplikácia metódy HTR+ (automatické rozpoznanie rukopisných textov) na historické rukopisy slovenskej proveniencie

Imrich Nagy

imrich.nagy@umb.sk

Táto prezentácia je výstupom z riešenia projektu APVV-19-0456
SKRIPTOR – Inovatívne sprístupnenie písomného dedičstva
Slovenska prostredníctvom systému automatickej transkripcie
historických rukopisov

Handwritten Text Recognition (HTR+)

- metóda na rozpoznávanie a automatický prepis historických rukopisov využívajúca technológie založené na neural engines, ktorá bola vyvinutá v rámci projektu Horizon 2020 Recognition and Enrichment of Archival Documents (READ), G. Mühlberger (Universität Innsbruck), 2016 – 2019: https://cordis.europa.eu/project/id/674943
- platforma Transkribus: https://readcoop.eu/transkribus/?sc=Transkribus
- na Slovensku ako prvý upozornil na pozoruhodné výsledky projektu aj s analýzou možností ich aplikácie a využitia v našich podmienkach prof. Katuščák. (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)
- rozvoj Transkribusu: združenie READ-COOP SCE, ktoré má v súčasnosti už 86 členov z 24 krajín https://readcoop.eu/members/

Projekt Skriptor

Cieľ:

- overiť využiteľnosť nástroja *Transkribus* pri sprístupnení obsahu digitalizovaného rukopisného materiálu rôznorodej proveniencie i datovania

Spolupráca s pamäťovými inštitúciami

Odbor archívov a registratúr MV SR

- Slovenský národný archív, Bratislava, Archív rodu Zay – Bučiansky archív, zbierka Martin Lauček: Collectanea, 2. pol. 18. storočia
- SNA, špecializované pracovisko Slovenský banský archív v Banskej Štiavnici, Banská komora v Banskej Bystrici, Metales (Metácia chotára mesta Banská Bystrica so 7 mapami), 1820 – 1823
- Štátny archív v Banskej Bystrici, fond Koháry – Coburg (1241) 1321 – 1945, časť V., číselný katalóg korešpondencie, 1. pol. 20. storočia

Spolupráca s pamäťovými inštitúciami

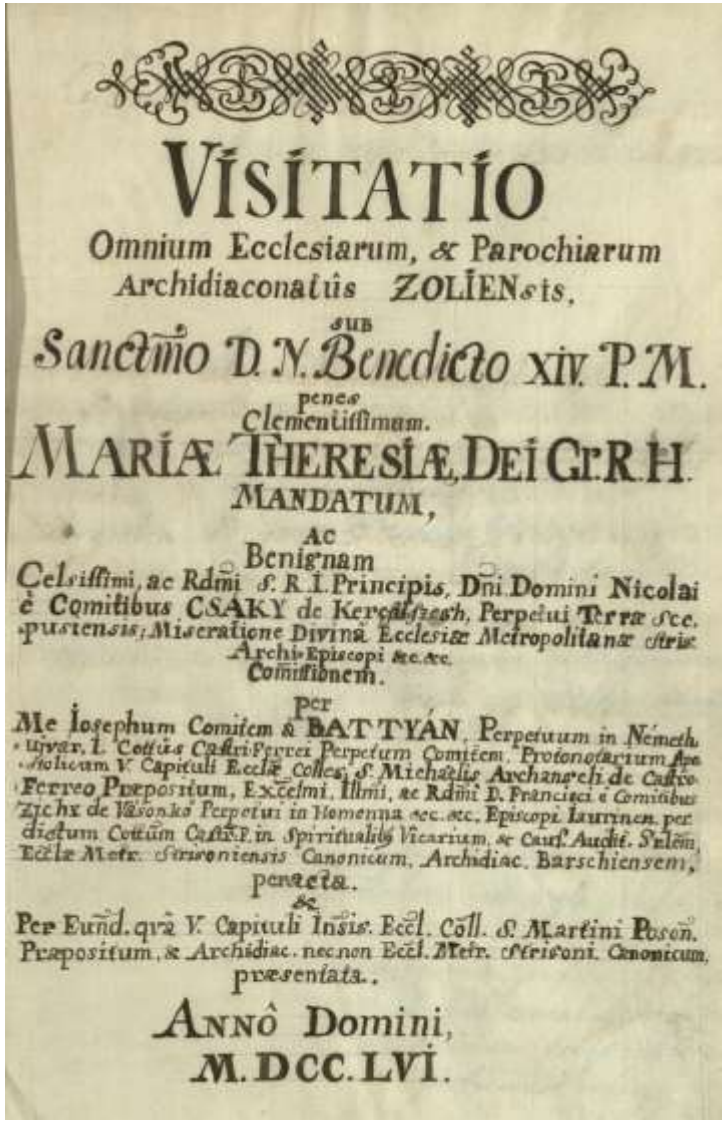
Slovenská národná knižnica, Martin, Literárny archív

- Náučné a iné práce, III-C, C 1069, Lauček, Martin: Collectanea, 2. pol. 18. storočia
- Literárne rukopisy, XXXI, 240 BN 1, Abrahamides Hrochotský, Izák: Postila, 1600 – 1601
- Zbierka korešpondencie, VIII, fond Hurbanovci, Jozef Miloslav Hurban (korešpondencia odoslaná J. M. Hurbanom), 1838 – 1887

Spolupráca s pamäťovými inštitúciami

Rímskokatolícka cirkev, Biskupstvo Banská Bystrica, Diecézny archív

- Kanonická vizitácia Zvolenského arcidiakonátu, 1756



VISITATIO

Omnium Ecclesiarum, & Parochiarum Archidiaconatûs ZOLIENsis,

sub

Sanctmo D. N. Benedicto XIV. P. M.

penes

Clementissimam.

MARIÆ THERESIÆ, DEI Gr. R. H.

MANDATUM,

Ac

Benignam

Celsissimi, ac Rdmi S. R. I. Principis, Dni Domini Nicolai è Comitibus CSAKY de Kerecsseth, Perpetui Terræ &c. pusiensis; Miseratione Divinâ Ecclesiæ Metropolitanæ Strig. Archi-Episcopi &c. &c.

Comissionem.

Per

Me Iosephum Comitem à BATTYÁN, Perpetuum in Németh-Ujvar, L. Cottûs Castri Ferrei Perpetuum Comitem, Protonotarium Apostolicum, V. Capituli Eccliæ Colleg. S. Michaëlis Archangeli de Castro Ferreo Præpositum, Excelmi, Illmi, ac Rdmi D. Francisci è Comitibus Zichy de Vásonkö Perpetui in Homonna &c. &c. Episcopi Iaurinen. per dictum Coetum Cathol. in Spiritualib. Vicarium, & Cauſ. Auditt. Solen. Eccliæ Metr. Strigoniensis Canonicum, Archidiac. Barschiensem, peracta.

&

Per Eund. qvâ V. Capituli Insig. Eccl. Coll. S. Martini Posoniensis Præpositum, & Archidiac. necnon Eccl. Metr. Strigoni. Canonicum, præsentata.

ANNÔ Domini,

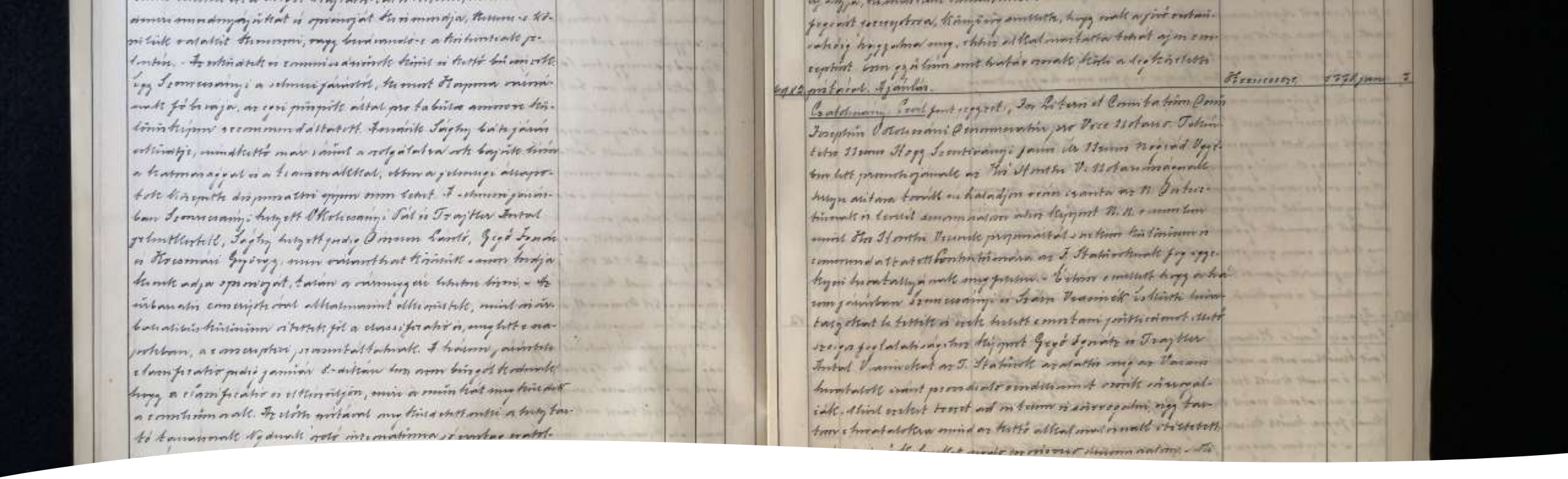
M. DCC. LVI.

Automatická transkripcia historického rukopisu

1. digitalizácia rukopisu a jeho nahratie do Transkribusu
2. segmentácia
3. tvorba modelu automatickej transkripcie a jeho aplikácia
4. výstupy z Transkribusu

Digitalizácia historického rukopisu

- digitalizáty z pamäťových inštitúcií v rozlíšení 600, 400, 300 dpi
- digitalizácia pomocou smartfónov (s využitím ScanTent-u a vhodnej mobilnej aplikácie – napr. DocScan pre Android)
- https://readcoop.eu/scantent/
- - použité smartfóny:
    - Apple Iphone 11 Pro
    - Samsung Galaxy S20+
    - Huawei Pro 40
    - Google Pixel 4



Digitalizát snímaný cez Iphone+ScanTent (priemerné dpi 192), súborový formát JPEG:

dobová archívna pomôcka – číselný katalóg korešpondencie z archívneho fondu Koháry – Coburg, ktorý spracoval bratislavský archivár János József Csákós v rokoch 1944 – 1945. Katalóg obsahuje regesty 6632 listov rôzneho rozsahu (v niektorých prípadoch obsahujúce aj výťahy, resp. preklady originálnych listov in extenso) v maďarskom jazyku spísané v rukopise (moderné kurzívne písmo) v podobe tabuliek na 4140 stranách formátu A3 (250 x 400 mm) v zošitovej väzbe

Transkribus podporuje formáty: PDF, JPEG, PNG a TIFF

Segmentácia rukopisu (určenie štruktúry a orientácie textu)

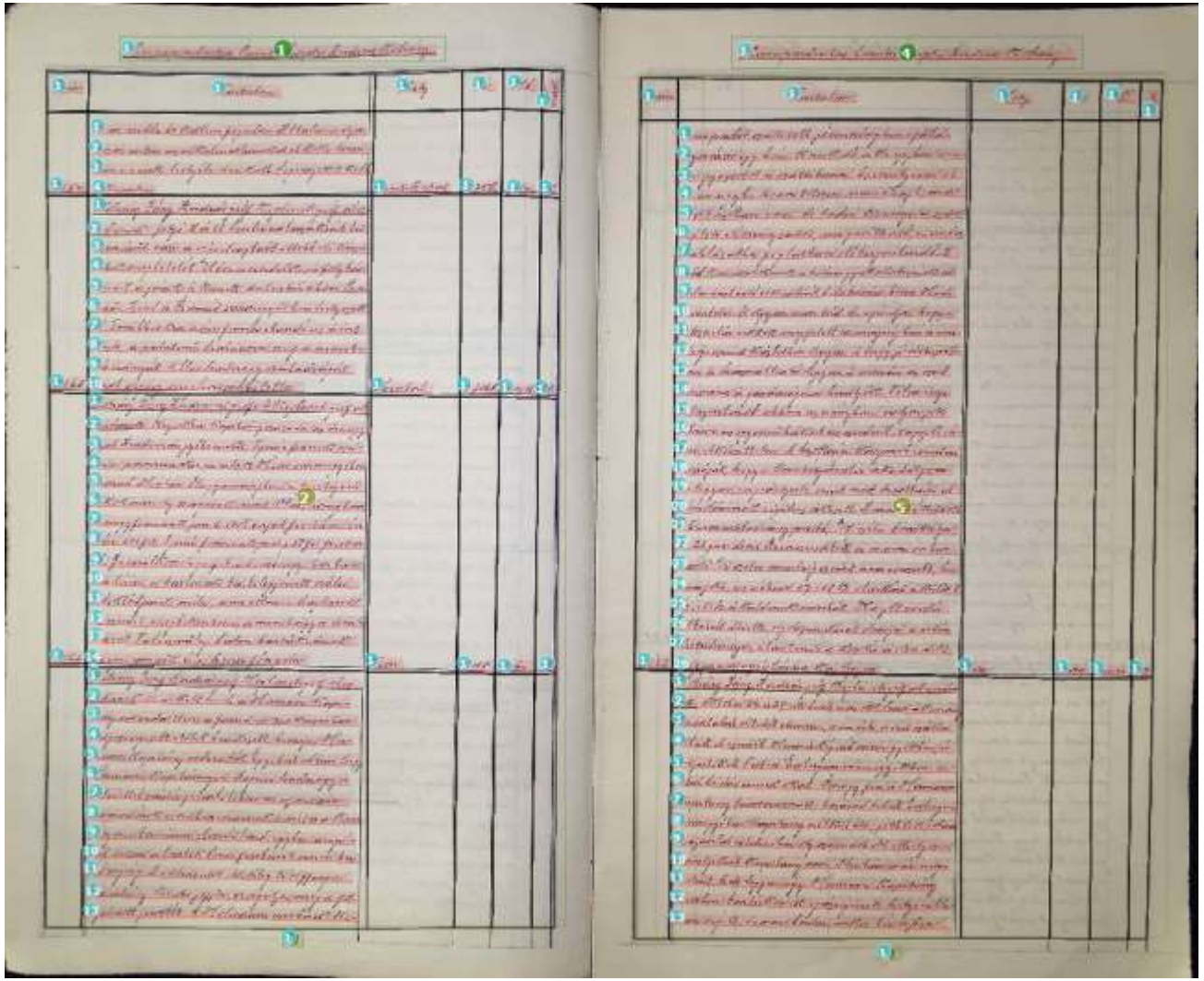
- automatická
- manuálna (tabuľky)

→ určenie textových rámcov

→ určenie hraníc riadkov

→ určenie poradia čítania

→ časová náročnosť úplnej segmentácie jednej dvojstrany bola priemerne 10 minút

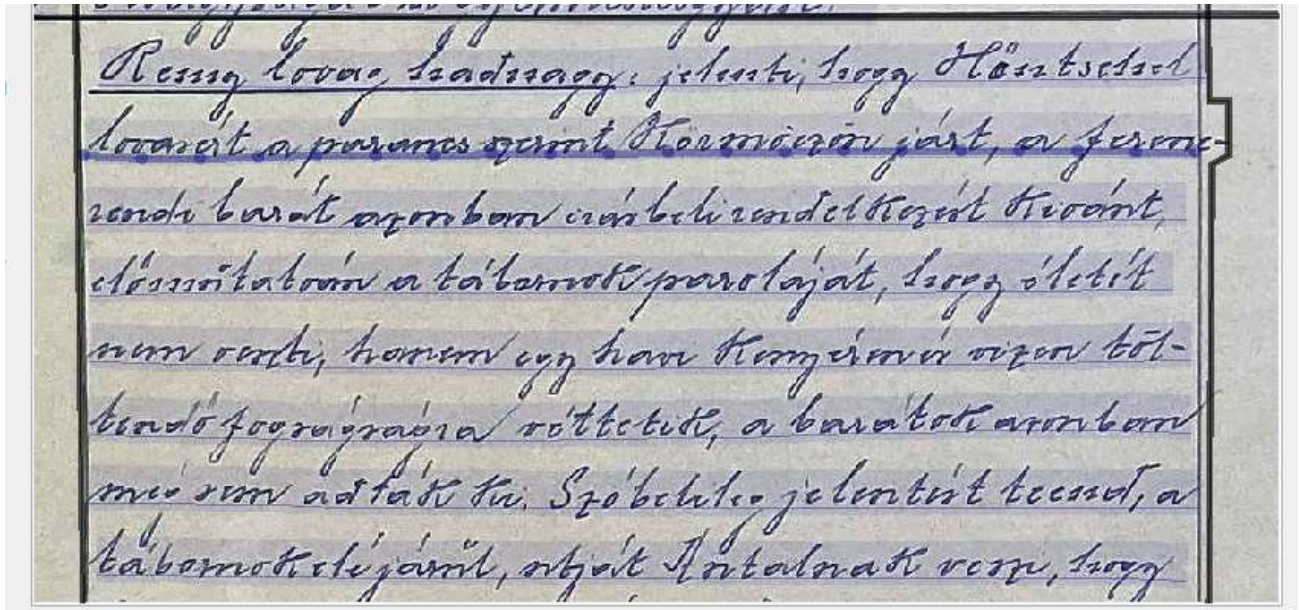


Príprava tvorby modelu automatickej transkripcie

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.

V našom prípade išlo o 29 snímok, čiže digitálnych obrazov obsahujúcich prvých 53 strán zo 4140 stranového rukopisu.

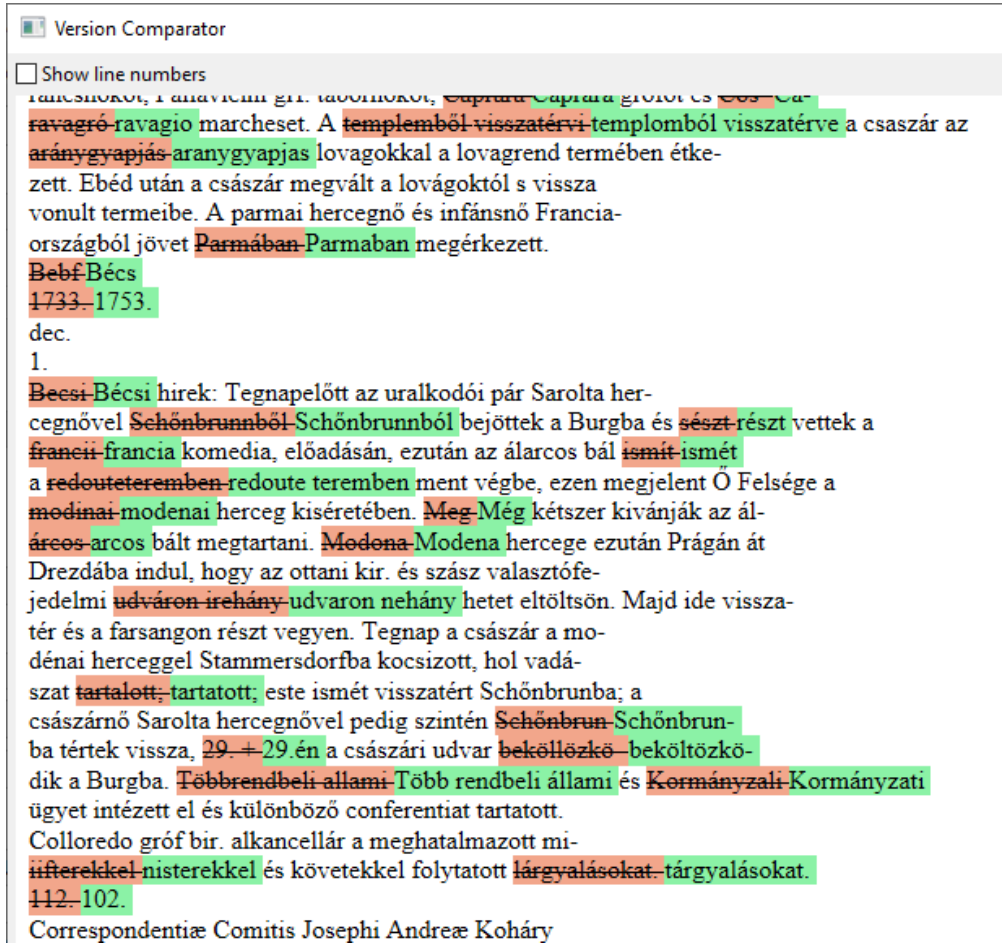


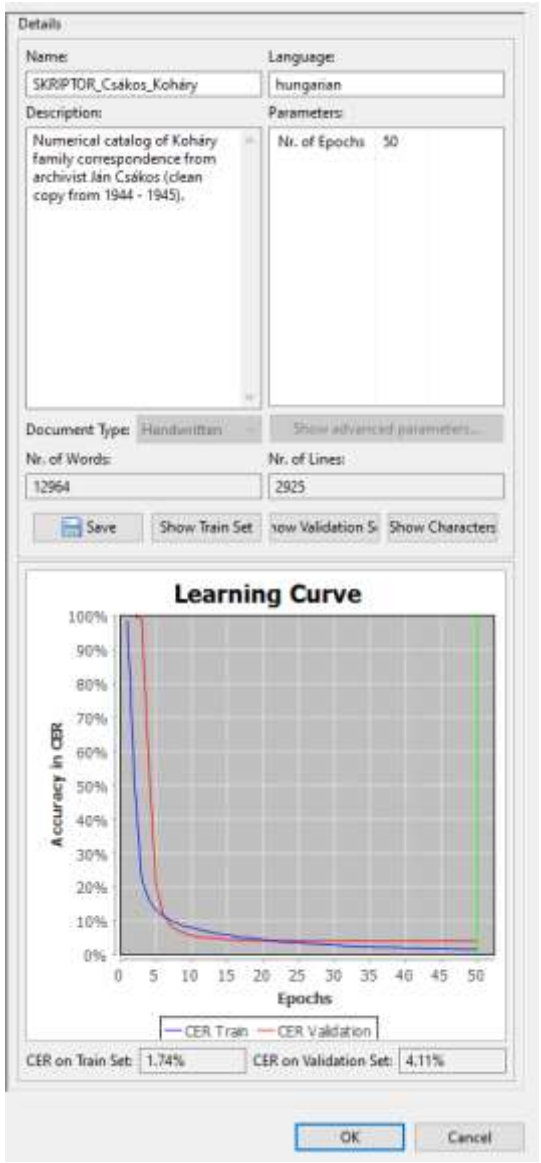
33-1 Remy· lovag· hadnagy:· jelenti,· hogy· Köntschel
33-2 lovasért· a· parancs· szerint· Körmöczön· járt,· a· ferenc-
33-3 rendi· barát· azonban· irásbeli· rendelkezést· kivánt,
33-4 előmutatóan· a· tábornok· paroláját,· hogy· életét
33-5 nem· veszti,· hanem· egy· havi· kenyéren· és· vizen· töl-
33-6 tendő· fogságságra· véttetik,· a· barátok· azonban
33-7 mégsem· adták· ki.· Szóbelileg· jelentést· teend,· a
33-8 tábornok· elé· járul,· utját· Antalnak· veszi,· hogy

Tvorba modelu automatickej transkripcie

Prepísaná vzorka rukopisu (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). 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. Počet cyklov si používateľ môže nastaviť aj na inú hodnotu 100, 200, 1000..., čo môže mať vplyv na spresnenie modelu.

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.





Hodnotenie modelu automatickej transkripcie

V charakteristike modelu máme k dispozícii 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).

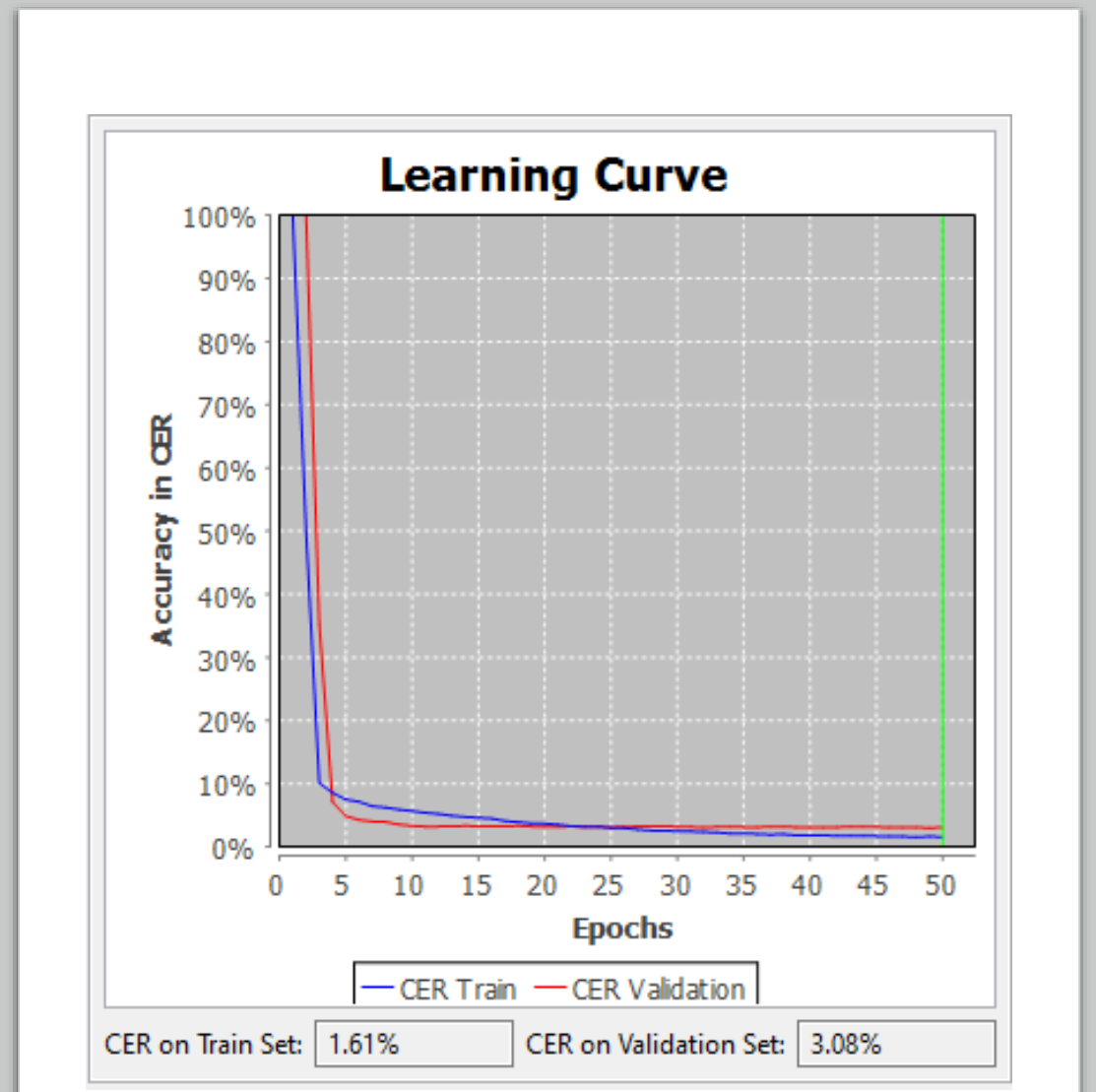
Aplikácia modelu automatickej transkripcie

Overenie modelu na ďalšej dávke 28 snímok, ktoré obsahovali strany 54 – 105 Csákósovho rukopisu:

- najprv štandardná príprava digitalizátu rukopisu v Transkribuse (segmentácia všetkých 28 strán až na úroveň určenia hraníc riadkov a kontroly ich poradia)
- proces automatickej transkripcie založený na vytrénovanom modeli

Vyhodnotenie úspešnosti automatickej transkripcie:

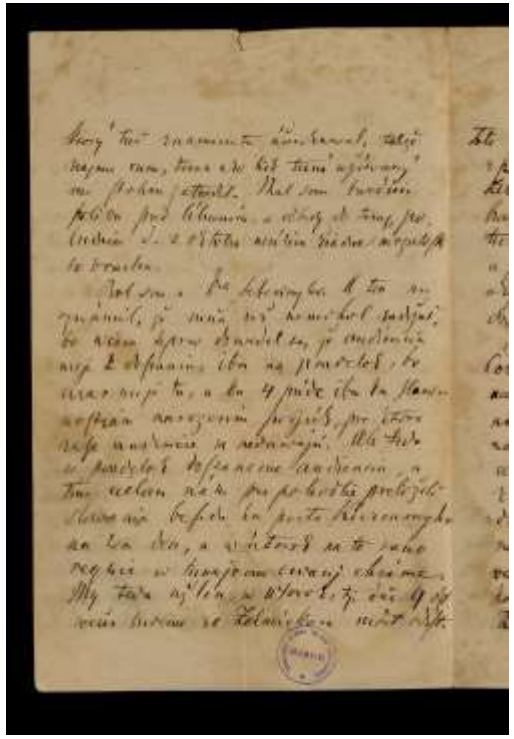
- porovnanie s manuálne korigovaným textom príslušných strán rukopisu (Ground Truth)
- z výsledkov vyplýva, že CER sa pohybuje v rozmedzí od 1,67 % (excelentný výsledok) po 8,12 % (použiteľný výsledok). Priemer CER zo všetkých porovnávaných strán 5,26 % s malou odchýlkou (1,15 percentuálneho bodu) zodpovedá CER nášho modelu 4,11 %
- možnosť zdokonalenia pôvodného modelu novou vzorkou Ground Truth:
- 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).

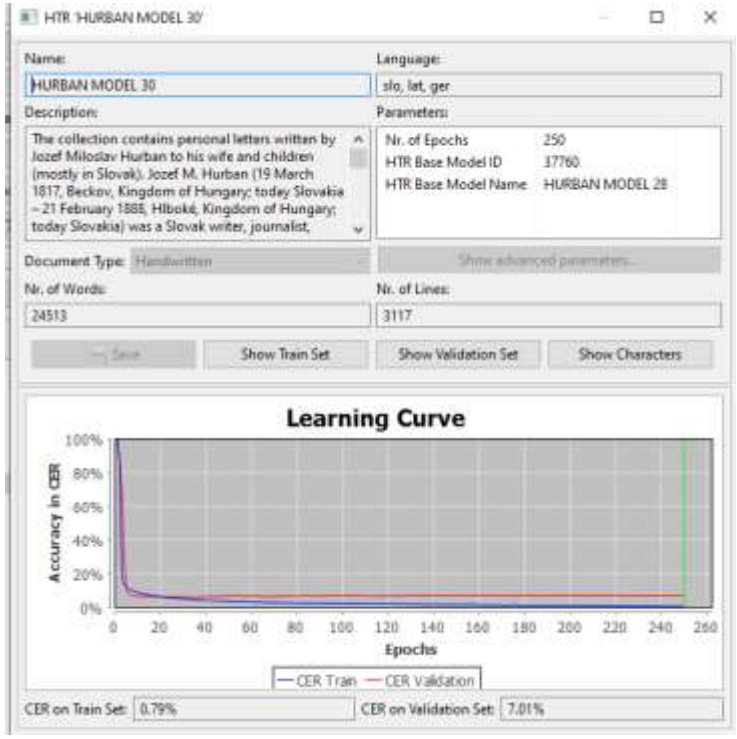


Výsledky modelu automatickej transkripcie pre rukopis J. M. Hurbana

(spracovala dr. A. Kurhajcová)

- digitalizát 600 dpi
- použitá vzorka 139 strán (123 pre training set, 16 pre validation set)
- počet cyklov trénovania modelu
- CER on Validation Set: 7,01 %





Výstupy z Transkribusu

digitálny text, ktorý možno 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)

Szám	Tartalom	Hely	Év	Hó	Nap.
2159.	it az ezredbe be kellene fogadni. Altalános ujon czozás esetén az alkalmatlanokat el kelle bocsátani s ezek helyébe derekább legényeket kell fölavatni	Szent-Antal	1750.	jun.	8.
2160	Koháry József András gróf Keglevich gróf alezredesnek: folyó 8. és 10. levelei válaszaként tudomásul veszi a végrehajtást illető előkészületek megtételét. Udvari rendelkezés folytán ezredét a jászok és kúnok districtusában, Csongrád-, Arad-és Csanád vármegyékben helyezik el. Továbbiakra nézve fenntartandó az érintkezés: a palatinus határozza meg a menetelés irányát. Allio hadnagy szabadságát szept. végeig meghosszabbította.	Ebental	1748.	szept.	18.

Prezentácia výsledkov transkripcie

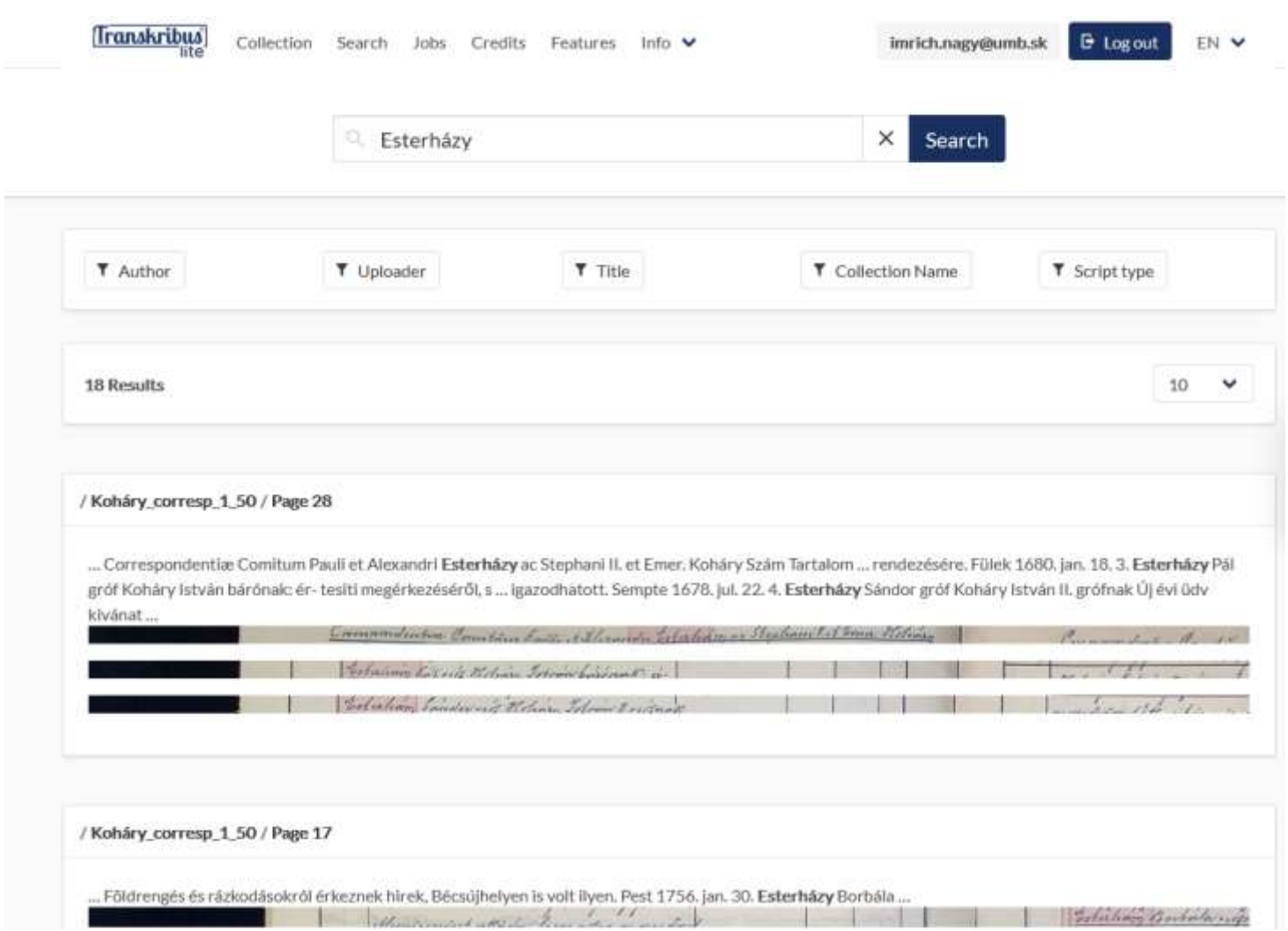
Pre sprístupnenie ponúka vhodný nástroj samotný Transkribus. Ide o webové rozhranie „read&search“, ktorý ponúka prezeranie digitalizovaného originálu paralelne s jeho prepisom s graficky zvýraznenou konkordanciou na úrovni riadkov.

https://readcoop.eu/readsearch/

Možnosti vyhľadávania v read&search

Objektom vyhľadávania sú samostatné slová, ktoré možno upraviť zástupnými znakmi pre rozšírenie vyhľadávania aj na alternatívne tvary slova:

Napr. 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.).



Záver

- 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
- metóda HTR+ pomocou platformy Transkribus ponúka účinný a účelný nástroj pre dosiahnutie tohto cieľa
- privítame záujem o spoluprácu pri tomto aplikačnom výskume