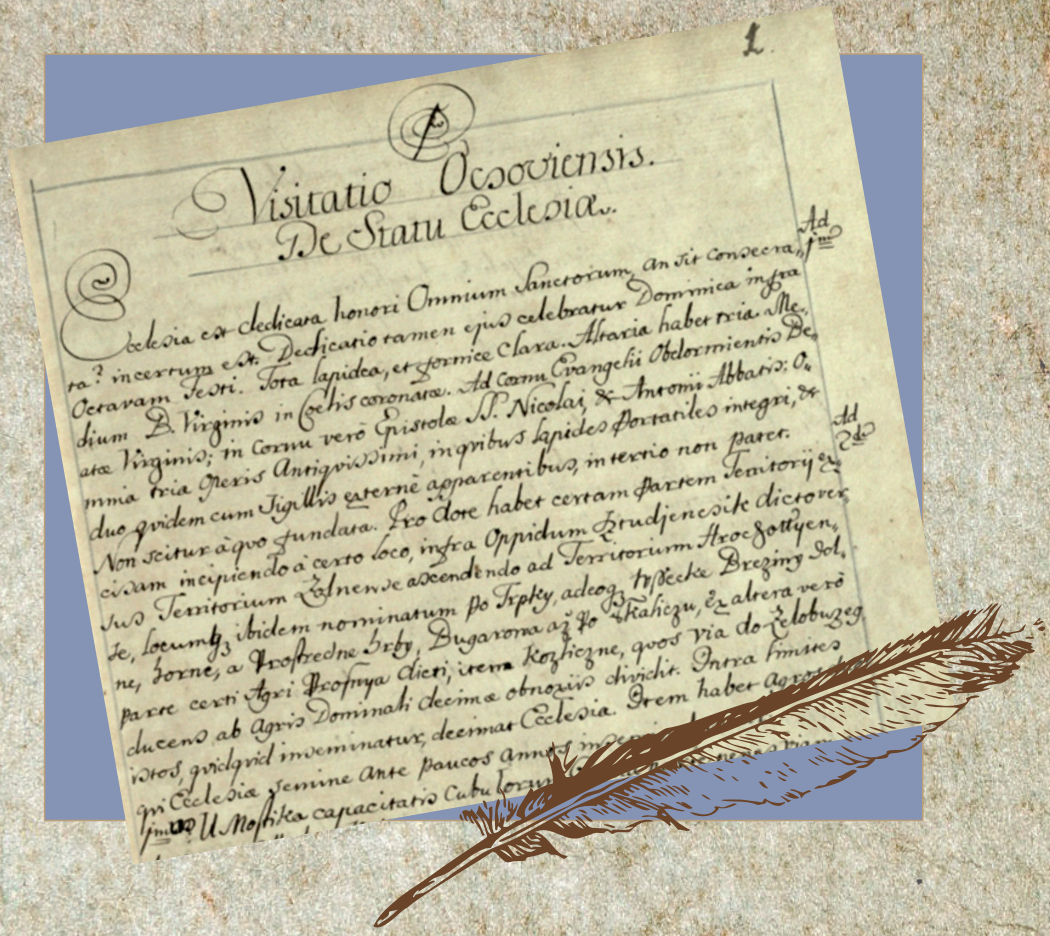


Dušan Katuščák – Imrich Nagy (eds.)

AUTOMATICKÁ TRANSKRIPCIA SLOVACIKÁLNYCH HISTORICKÝCH DOKUMENTOV

Belianum 2022

Dušan Katuščák – Imrich Nagy (eds.)

AUTOMATICKÁ TRANSKRIPCIA SLOVACIKÁLNYCH HISTORICKÝCH DOKUMENTOV

Prvé vydanie
Elektronická publikácia

BELIANUM
2022

KATUŠČÁK, Dušan a Imrich NAGY, eds.

Automatická transkripcia slovacikálnych historických dokumentov [elektronická editovaná kniha] / zost. Dušan Katuščák, Imrich Nagy ; rec. Milan Konvit, Jan Odstrčilík. - 1. vyd. - Banská Bystrica : Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2022. - 207 s. - ISBN 978-80-557-2020-3. - DOI https://doi.org/10.24040/2022.9788055720203

Elektronická publikácia je výstupom z riešenia projektu APVV-19-0456 SKRIPTOR – *Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov* (2020 – 2024).

Autori:	© Mária Bôbová; Dušan Katuščák; Alica Kurhajcová; Pavol Maliniak; Michaela Mikušková; Imrich Nagy; Lucia Nižníková; Patrik Kunec; Oto Tomeček
Recenzenti:	prof. Ing. Milan Konvit, PhD., Jan Odstrčilík, Ph.D.
Anglický preklad textov:	Mgr. Róbert Címer, Mgr. Zuzana Hušlová, Bc. Mária Onderufová
Spolupráca:	Ing. Ivana Poláková, PhD.
Jazyková korektúra:	PaedDr. Ivan Očenáš, PhD.
Grafická úprava:	PaedDr. Dušan Jarina

Táto práca bola podporená Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-19-0456 *SKRIPTOR – Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov*.

This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-19-0456 *SKRIPTOR – Innovative access to the written heritage of Slovakia through a system of automatic transcription of historical manuscripts.*

DOI: https://doi.org/10.24040/2022.9788055720203

ISBN 978-80-557-2020-3

ABSTRAKT

Editovaná vedecká kniha je čiastkovým výstupom z riešenia projektu APVV-19-0456 SKRIPTOR – *Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov* (2020 – 2024) po prvých rokoch jeho riešenia. Ide o kolektívne vedecké dielo, ktoré je zamerané na jednu tému, ktorou je *transkripcia historického písomného dedičstva*. Všetci autori jednotlivých kapitol sú riešiteľmi projektu. Kniha je koncipovaná ako súbor autorských prehľadových, overovacích a experimentálnych štúdií v samostatných kapitolách, v ktorých autori prinášajú historické výklady, interpretácie, ako aj nové poznatky, zistenia a skúsenosti týkajúce sa zbierok a dokumentov, ktoré sú predmetom výskumného projektu. Kniha prináša výsledky skúmania možností automatickej transkripcie historického písomného dedičstva prevažne jazykových, geografických a autorských slovacík. Cieľom skúmania bolo overiť možnosti použitia platformy *Transkribus*. Dielo obsahuje podrobne spracovaný vedecký aparát, zahŕňajúci aj bibliograficko-informačný, ilustračný, poznámkový a kritický aparát v autorských kapitolách.

ABSTRACT

The edited scientific book is a partial output from the solution of the project APVV-19-0456 SKRIPTOR – *Innovative access to the written heritage of Slovakia through the system of automatic transcription of historical manuscripts* (2020-2024) after the first years of the project solution. It is a collective scientific work that focuses on one theme, which is *the transcription of historical written heritage*. All authors of individual chapters are project investigators. The book is conceived as a set of authorial review, verification and experimental studies, in separate chapters in which the authors bring historical explanations, interpretations, as well as new knowledge, findings and experiences related to the collections and documents that are the subject of the research project. The book presents the results of exploring the possibilities of automatic transcription of the historical written heritage of predominantly linguistic, geographical and authorial Slavonics. The aim of the investigation was to verify the possibilities of using the *Transcribus* platform. The work contains a detailed scientific apparatus, including bibliographic-informational, illustrative, note-taking and critical apparatus in the author's chapters.

Predhovor

Posledné dve desaťročia boli obdobím realizácie projektov masovej digitalizácie na medzinárodnej a národnej úrovni, ako aj na úrovni rôznych informačných inštitúcií, knižníc, archívov, galérií a audiovizuálnych pracovísk. Výsledkom projektov digitalizácie sú mohutné repozitáre textových, obrazových, zvukových a audiovizuálnych digitálnych dokumentov.

Obsah knihy je zameraný na textové dokumenty, najmä na rukopisné dokumenty novoveku, ako aj na staré a vzácne tlače. V digitálnych repozitároch sú historické tlače s určitou presnosťou rozpoznané technológiou optického rozpoznávania. Avšak rukopisy sú dostupné používateľom väčšinou len ako nasnímané obrazy. Technologický pokrok a stále zdokonaľované aplikácie umelej inteligencie už dnes umožňujú na dobrej až excelentnej úrovni rozpoznávať historické rukopisy a sprístupňovať unikátne archívne dokumenty verejnosti.

Špecifickým predmetom nášho záujmu preto sú historické rukopisy, možnosti ich automatickej transkripcie a sprístupnenia na vedecké, výskumné a vzdelávacie ciele.

Víziou vedcov, expertov a iných používateľov z oblasti písomného dedičstva je, aby sa verejne dostupné modely transkripcie postupne stali užitočným spoločným nástrojom na automatickú transkripciu historických dokumentov. Je potrebné dosiahnuť takú úroveň, aby už nebolo potrebné tvoriť pre každú zbierku rukopisov a tlačí samostatný model. Pre používateľov by malo ísť o akúsi „čiernu skrinku“ (black box), v ktorej robot, umelá inteligencia sama vyberie z agregovaných 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 je nevyhnutné trpezlivo tvoriť množstvo parciálnych modelov.

Téme sa venuje náš projekt SKRIPTOR. Jeho zmyslom je, aby súčasťou spoločného medzinárodného úsilia boli aj naši odborníci. Usilujeme sa o to, aby budúca „čierna skrinka“ bola pripravená poskytnúť pomoc všetkým pri transkripcii unikátnych historických zbierok a dokumentov najmä v jazykoch nášho regiónu, medzi ktoré patrí slovenčina, čeština, poľština, lužická srbčina, latinčina, maďarčina. V súčasnej fáze vývoja sa sústreďujeme na zvládnutie nástrojov transkripcie a na tvorbu modelov transkripcie na základe väčších zbierok, ktoré obsahujú stovky a tisíce strán.

Kniha nadväzuje na vedeckú konferenciu *Digital humanities – nástroje sprístupňovania historického dedičstva* konanú v Banskej Bystrici v dňoch 12. – 13. 10. 2022. Táto kniha, podobne ako samotná konferencia a zborník abstraktov vydaný pre účastníkov konferencie, sú plánovanými výstupmi projektu SKRIPTOR.[11] Ide o projekt APVV-19-0456 SKRIPTOR (2020 – 2024) s názvom *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].* Jednou z aktuálnych recenzovaných publikácií je štúdia D. Katuščáka, ktorá má podobný obsah ako ostatné štúdie v tejto knihe, ale

1 MALINIAK, Pavol – NAGY, Imrich (eds.): *Digital humanities: nástroje sprístupňovania historického dedičstva. Zborník abstraktov.* Banská Bystrica : Štátna vedecká knižnica, 2022. 73 s.

nezaradili sme ju do knihy, pretože bola už publikovaná.[2]

Všeobecný kontextový rámec projektu SKRIPTOR a výskumu transkripcie historických dokumentov s použitím umelej inteligencie tvorí oblasť nazývaná *digital humanities*. Podľa nášho názoru možno *digital humanities* (digitálnu humanistiku) považovať za spoločné pomenovanie a prierezovú metodológiu pre všetky aplikácie informačných a komunikačných technológií (IKT) v spoločenských a humanitných vedách, odboroch a disciplínach a v im zodpovedajúcej praxi. Táto metodológia sa komplexne uplatnila v európskom projekte základného výskumu *READ*, ktorý sa realizoval v rámci programu *Horizon 2020*.

Priame podnety na zameranie slovenského projektu SKRIPTOR nám poskytli práve poznatky a nástroje európskeho projektu výskumu READ *Recognition and Enrichment of Archival Documents*.[3] Autorom a koordinátorom projektu bol prof. Günter Mühlberger z Univerzity v Insbrucku. Udržateľnosť projektu READ a aplikácia výsledkov je zabezpečená v združení READ-COOP (*A European Cooperative Society*). Združenie malo v roku 2022 vyše 100 členov z 27 krajín.

Kľúčový inovatívny nástroj na transkripciu historických rukopisných dokumentov je *Transkribus.* Je to 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. Existujú dva systémy platformy: prvý je *Transkribus Lite, ktorý* je možné použiť v prehliadači osobných počítačov a smartfónov. Druhý je *Transkribus Expert*, pričom mnohé jeho funkcie môžu byť použité aj v *Transkribus Lite*. Výsledky transkripcie sú dostupné cez portál *Read&Search*. Platforma *Transkribus* 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* v Rostocku. V súčasnosti s platformou pracujú tisíce historikov, archivárov, knihovníkov. Platforma bola vytvorená v kontexte dvoch predchádzajúcich projektov EÚ *tranScriptorium* (2013 – 2015) a READ (2016 – 2019).

Na Slovensku sme začali pracovať s platformou *Transkribus* v roku 2017 a informovali sme verejnosť o našich prvých modeloch transkripcie rukopisov. Spočiatku išlo o individuálnu iniciatívu, ktorá vďaka osvietenej ústretovosti ľudí z Univerzity Mateja Bela prerástla do inštitucionálneho výskumu v projekte SKRIPTOR.

V platforme *Transkribus* sme používali stroj umelej inteligencie *HTR+* (Handwritten Text Recognition) a *PyLaia*. Tieto stroje zatiaľ nemôžu okamžite automaticky tran-

2 KATUŠČÁK, Dušan: Umelá inteligencia pomáha sprístupňovať písomné dedičstvo. In: *Knihovna : knihovnická revue* [online], roč. 33, č. 2, 2022, s. 27 [cit. 2022-11-20]. Dostupné na: https://knihovnarevue.nkp.cz/archiv/2022-2/recenzovane-prispevky/umela-inteligencia-pomaha-spristupnovat--pisomne-dedicstvo

3 MÜHLBERGER, Günter: *READ (Recognition and Enrichment of Archival Documents) – 2016–2019* [online]. [cit 2021-10-06.] Dostupné na: https://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019
MUEHLBERGER, Guenter et al.: Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: *Journal of Documentation* [online], vol. 75, no. 5, 2019, pp. 954 – 976 [cit. 2021-10-06]. Dostupné na: https://doi.org/10.1108/JD-07-2018-0114

skribovať rôzne historické rukopisy. Najprv musí byť stroj vyškolený na konkrétny typ písma a rukopisu. Hlavným cieľom praktických experimentov v projekte SKRIPTOR a predstavených v tejto monografii je v súčasnosti tvorba *modelov* transkripcie.

Výskumníci zapojení do projektu SKRIPTOR po počiatočnej nedôvere k možnostiam transkripcie sa postupne stávajú expertmi. Príspevky v monografii sú dokladom toho, že si osvojujú nástroj *Transkribus*, že zvládli základné pracovné postupy a že čoraz dôkladnejšie spoznávajú funkcionalitu platformy *Transkribus*. Darí sa im vytvárať veľmi dobré až excelentné modely transkripcie archívnych dokumentov a starých tlačí.

Niektorí predstavitelia *digital humanities* na Slovensku majú k tejto iniciatíve – ako ku podozrivej novote – 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“. Je to známy jav aj z iných zápasov tradicionalistov a novátorov. Často ide o reakcie, ktoré na jednej strane síce verbálne deklarujú záujem o „digitalizáciu“ a „umelú inteligenciu“, no na druhej strane tento záujem ďalej nepokračuje získavaním potrebných vedomostí a hlavne nesvedčí o dostatočných vedomostiach o problematike a možnostiach digitalizácie a využitia umelej inteligencie aj pre tradičné humanitné odbory. Problémom je zrejme aj fakt, že *Transkribus* nie je hotový nástroj, „policový softvér“ hotový na „klikanie“, ale nástroj, ktorý sa kolektívne stále tvorí a zdokonaľuje.

Parciálne štúdie prinášajú výsledky a modely, ktoré môžu byť použité vo väčších agregovaných modeloch READ-COOP. Uvedomujeme si, že modely transkripcie špecifických dokumentov našej kultúrnej proveniencie, ktoré sú súčasťou európskeho písomného dedičstva, za nás nikto neurobí. Postoje niektorých 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. Veríme, že predkladaná monografia povzbudí mnohých podporiť trend sprístupňovania historického písomného dedičstva inovatívnymi nástrojmi a formami.

Od začiatku roka 2023 sa budú výskumníci zoznamovať s platformou Transkribus Lite, ktorá sa neustále vyvíja a prenáša do webového prostredia. Členovia beta programu Transkribus dostávajú po prihlásení automaticky nové verzie a funkcie Transkribus Lite skôr, ako sú dostupné širokej verejnosti. To umožňuje vyskúšať si najnovšie vylepšenia a poskytnúť spätnú väzbu, ktorá pomôže formovať budúci vývoj Transkribus Lite, zatiaľ s limitom veľkosti nahratého súboru 10 MB.

Tím Transkribus v súčasnosti aktualizuje niekoľko zastaraných komponentov Transkribu. Najnovším komponentom je funkcia analýzy rozloženia, ktorá zisťuje čiary v každom obrázku pred spustením rozpoznávania textu, alebo ktorú je možné spustiť nezávisle od rozpoznávania textu. Nový variant nahrádza CITLab Advanced a možno ho použiť výberom Transkribus LA v príslušnom dialógu. Ak ju pred spustením rozpoznávania textu nechceme spúšťať samostatne, je predvolene vybratá vždy, keď sa spustí úloha rozpoznávania textu a nemusíme sa o ňu starať.

Všetky nové úlohy analýzy rozloženia, ktoré sú odteraz spustené s vybratou CITlab Advanced, budú spracované novým komponentom Transkribus LA a s predvolenými nastaveniami.

Úlohou riešiteľov projektu SKRIPTOR je tvoriť modely, ktoré umožnia transkripciu písomného dedičstva z našej kultúrnej a jazykovej oblasti, pre ktorú sú charakteristické určité druhy písma, jazyky, znaky, štýly, diakritika ap. Výskum bude pokračovať pretrénovaním modelov preferovaným strojom *PyLaia* a tvorbou niekoľkých agregovaných, základných modelov na transkripciu historických dokumentov písaných v západoslovanských jazykoch, ako aj v iných jazykoch, ktoré sa historicky bežne vyskytovali v dokumentoch v našej kultúrnej oblasti, medzi ktoré patrí latinčina, nemčina, maďarčina. Rovnako sa pozornosť zameria na šírenie našich skúseností vo vzdelávaní a v komunite vedcov a odborníkov z archívov, knižníc a z akademickej sféry. Otvorená je zatiaľ otázka udržateľnosti projektu, ako aj otázka inštitucionalizácie a prepojenia so vzdelávaním a odbornou komunitou. Zaujímavou výzvou je možnosť prekladu historických textov získaných transkripciou zdokonaľovaním strojového prekladu.

December 2022

prof. PhDr. Dušan Katuščák, PhD.
doc. PhDr. Imrich Nagy, PhD.
editori

Obsah

Zoznam ilustrácií

Zoznam tabuliek

Zoznam skratiek a symbolov

A	Automatická segmentácia
ANN	Artificial Neural Network
CER	Character Error Rate
CITlab	Computational Intelligence Technology Lab
ERNiE	Encyklopedia of Romantic Nationalism in Europe
GT	Ground Truth
HTML	HyperText Markup Language
HTR	Handwritten Text Recognition
ISAD(G)	International Standard Archival Description (General)
JPG	Joint Photographic Experts Group
KWS	Keyword Spotting
LA SNK	Literárny archív Slovenskej národnej knižnice v Martine
LR	Line Regions
M	Manuálna segmentácia
M + A	Kombinovaná manuálna a automatická segmentácia
Model JMH	Model J. M. Hurban
OCR	Optical Character Recognition
PDF	Portable Document Format
PNG	Portable Network Graphics
RAW	Neupravený obrazový formát
READ	Recognition and Enrichment of Archival Documents
READ-COOP	A European Cooperative Society
RT HTR+	Retrained Handwritten Text Recognition
RTF	Rich Text Format
TIFF	Tag Image File Format
TR	Text Region
WER	Word Error Rate

KAPITOLA 1

METODOLÓGIA A METODIKA TRANSKRIPCIE HISTORICKÝCH TEXTOV

Dušan Katuščák

Slezská univerzita v Opavě; Filozoficko-přírodovědecká fakulta; Ústav bohemistiky a knihovnictví
E-mail: dusan.katuscak@fpf.slu.cz
Štátna vedecká knižnica v Banskej Bystrici
E-mail: dusan.katuscak@svkbb.eu

Abstrakt

Témou kapitoly je základný popis metodologického kontextu rozpoznávania textov historických dokumentov. Rozpoznávanie znakov v písomných dokumentoch autor považuje za oblasť vedeckej a praktickej činnosti, v ktorej sa používajú a ďalej vyvíjajú metódy, cieľavedomé postupy, nástroje, operácie, pravidlá a súbory pravidiel. Z používateľského hľadiska stručne opisuje OCR a HTR ako metódy rozpoznávania. Definuje základné pracovné postupy transkripcie na základe vlastnej empírie a experimentov v rámci aplikovaného výskumu *SKRIPTOR*. Vysvetľuje pojmy, termíny, postupy, metódy a nástroje transkripcie podľa hlavných procesov transkripcie: príprava, inštalácia systému *Transkribus Expert Client*, snímanie, postprocesing, import dokumentov, segmentácia a manuálna transkripcia, tvorba modelu transkripcie a sprístupnenie a použitie výsledkov transkripcie.

Kľúčové slová: historické texty; metodológia; nástroje; OCR+; postupy; PyLaia; *Transkribus*; transkripcia

Abstract
Methodology and methodicalness of transcription of historical texts

The theme of this chapter is a basic description of the methodological context of text recognition of historical documents. He considers character recognition in written documents to be a field of scientific and practical activity in which methods, goal-oriented procedures, tools, procedures, operations, rules and sets of rules are used and further developed. Briefly describes OCR and HTR as recognition methods from a user perspective. It defines the basic workflows of transcription based on the author‘s own experience and experiments within the *SKRIPTOR* applied research. It explains concepts, terms, procedures, methods and tools of transcription according to the main processes of transcription: preparation, installation of the *Transkribus Expert Client* system,

scanning, postprocessing, document import, segmentation and manual transcription, creation of a transcription model and making available and using transcription results.

Key words: historical texts; methodology; OCR+; procedures; PyLaia; tools; *Transkribus*; transcription

ÚVOD

Rozpoznávanie znakov v písomných dokumentoch je oblasť vedeckej a praktickej činnosti, v ktorej sa určitým spôsobom používajú a ďalej vyvíjajú metódy, cieľavedomé postupy, nástroje, operácie, pravidlá a súbory pravidiel.

Oblasť transkripcie textov predstavuje v súčasnosti dynamický metodologický systém, ktorý je obohacovaný aj o metodologické mikrosystémy a know-how tvorcov modelov transkripcie.

V systéme historickej vedy sa transkripcia historických prameňov formuje ako nová metóda a prierezová disciplína pomocných vied historických. Automatická transkripcia významne rozširuje možnosti skúmania, poznávania a interpretácie dejín na základe masovejšie transkribovaných a všeobecne dostupných písomných rukopisných a tlačených prameňov úradného charakteru, ako aj prameňov súkromnej povahy, ako sú denníky, katalógy, korešpondencia, poznámky, zápisky, memoáre, anály, legendy a pod. Transkripcia historických prameňov má prierezový charakter vo vzťahu s inými disciplínami pomocných vied historických. Má technologickú podstatu a predpokladá úzku spoluprácu s inými špeciálnymi vednými disciplínami, ako je archivistika, diplomatika, kodikológia a paleografia.

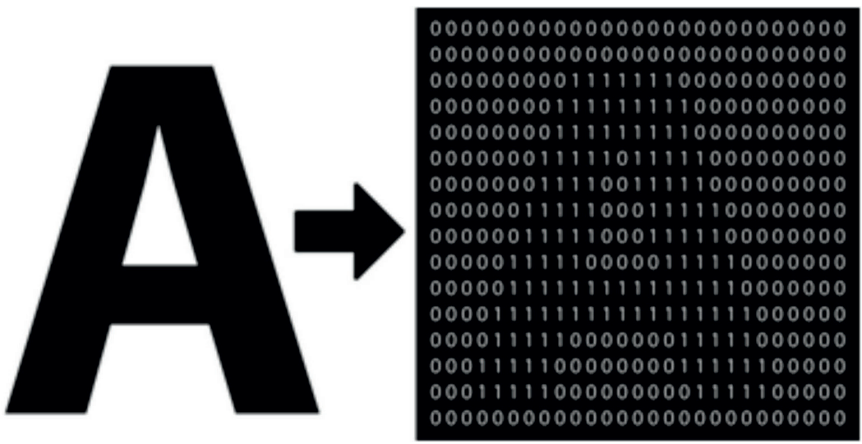
Automatická transkripcia predstavuje prechod od prevládajúceho parciálneho, fragmentovaného prístupu ku zbierkam a dokumentom k masovému sprístupneniu zbierok a dokumentov. Otvára možnosti systematického výskumu a sprístupnenia najvýznamnejších historických fondov, zbierok a dokumentov. Automatická transkripcia a digitalizácia archívnych a knižničných zbierok znamená zmenu paradigmy uplatnením *komunikačného* prístupu. Na rozdiel od *statického* zhromažďovania zbierok a dokumentov kladie dôraz na ich sprístupnenie a použitie.

Tento metodologický systém sa vyznačuje tým, že sa v ňom uplatňujú aj metódy a postupy rôznych iných vedných odborov a disciplín, ako je jazykoveda (transkripcia a transliterácia, morfológia, sémantika, lexikológia), informatika (umelá inteligencia), knižničná a informačná veda (bibliografia, digitalizácia).

Východiskovými danosťami uplatňovania určitých metód v tejto oblasti sú písomné dokumenty v analógovej a digitálnej forme. Cieľom uplatňovania týchto metód je špecifická hodnota, ktorú predstavuje *model* ako výsledok transkripcie. Výsledkom transkripcie je transkribovaný text, produkt, ktorý predstavuje v dokumentovej komunikácii pridanú hodnotu, nakoľko podstatne rozširuje možnosti využívania unikátnych historických textov z fondov pamäťových inštitúcií a poskytovanie digitálnych služieb.

DVE METÓDY ROZPOZNÁVANIA ZNAKOV – OCR A HTR

Teoretické alebo používateľské príspevky v oblasti rozpoznávania je možné rozdeliť do dvoch skupín podľa toho, či sa venujú rozpoznávaniu *tlačených textov* (OCR) alebo *rukopisných textov* (HTR). Informačné zdroje k téme OCR sa týkajú jednak pokračujúcich teoretických výskumov zameraných na samotnú umelú inteligenciu. Autormi týchto teoretických diel sú najmä *informatici a matematici*. Oni zohrávajú kľúčovú úlohu aj pri vývoji aplikácií OCR aj HTR. Optické rozoznávanie znakov OCR je metóda umožňujúca prevod obrazu (grafiky) tlačených alebo písaných znakov do textovej, editovateľnej formy. Podstata metódy OCR spočíva v použití softvéru a mapovaní každého znaku do zbierky pixelov, ktoré ho reprezentujú v určenom písme.



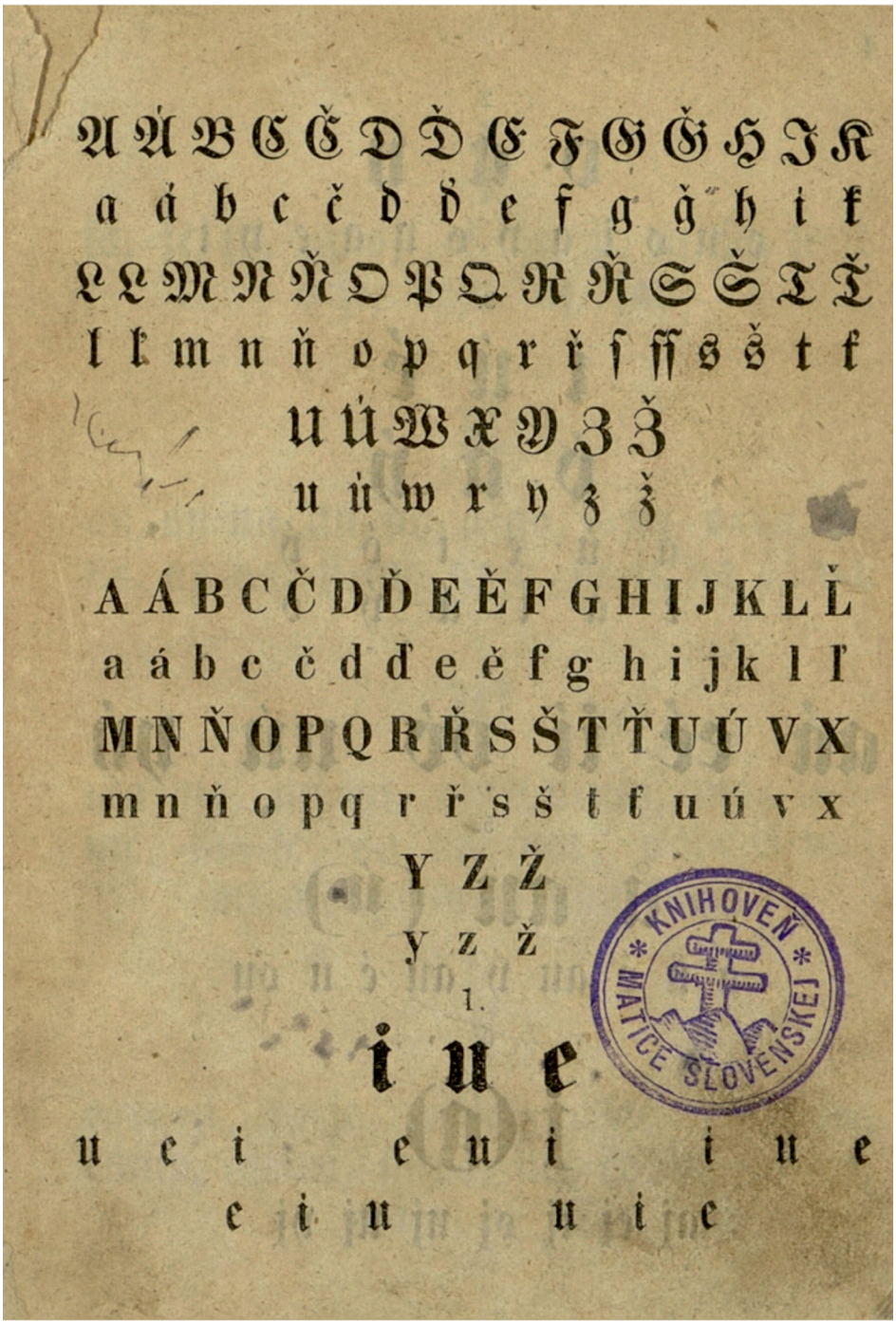
Obrázok 1 Prevod obrazu znaku písmena A do počítačovej formy na jednotky digitálnej rastrovej (bitmapovej) grafiky[1].

Metóda OCR umožňuje mimoriadne dobré rozpoznávanie tlačených a strojopisných znakov už od roku 1970. Existuje mnoho softvérov na OCR tlačených textov. K najvýznamnejším patrí *ABBYY FineReader*. Umožňuje konverziu snímok, napríklad fotografií, naskenovaných dokumentov, PDF súborov do upravovateľného digitálneho formátu do súborov typu Microsoft Word, Microsoft Excel, Microsoft PowerPoint, RTF, HTML, PDF, PDF/A, prehľadávateľné PDF, CSV, obyčajné textové súbory a pod. Verzia *ABBYY FineReader* 12 podporuje rozpoznávanie textu v 190 jazykoch a umožňuje kontrolu pravopisu pre 48 jazykov. Podstata tejto metódy spočíva v tom, že jednotlivé znaky, písmená alebo celé znakové sady sú naprogramované ručne.

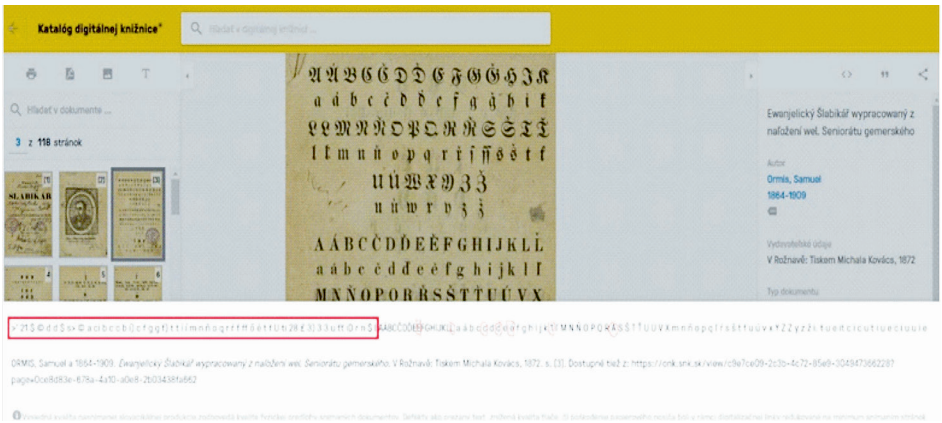
METÓDA OCR

V projektoch masovej digitalizácie tlačí sa používa nástroj na automatické rozpoznávanie – OCR starých tlačí Server. Pritom kvalita OCR býva sporná a pre rukopisy nevhodná.

1 DIETRICH, Felix: OCR vs. HTR or "What is AI, actually?" In: *READ-COOP* [online]. Innsburck : READ-COOP SCE, 2021 [cit. 2022-11-20]. Dostupné na: https://readcoop.eu/insights/ocr-vs-htr/



Obrázok 2 Abeceda v šlabikári z roku 1872. Fraktúra a antikva.



Obrázok 3 Automatické OCR abecedy v Ormisovom šlabikári z roku 1872.

Poznámka: V riadku pod obrázkom je zobrazený výsledok automatického OCR. Fraktúra v červenom rámčeku je rozpoznaná celkom neprijateľne a rovnako sú chyby v rozpoznávaní antikvy – červené značky. Zdroj: Slovenská národná knižnica: Katalóg digitálnej knižnice. Dostupné: Ewanjelický Šlabikář wypracowaný z naľožení wel. Seniorátu gemerského | Slovenská národná knižnica | Digitálna knižnica (snk.sk).

Podobné výsledky OCR ilustrujú ďalšie obrázky zo šlabikára z roku 1879.

Obrázok 4 Problematické automatické OCR tlačeného a rukopisného písma zo šlabikára z roku 1879. Zdroj: Slovenská národná knižnica: Katalóg digitálnej knižnice. Dostupné: Ewanjelický Šlabikář wypracowaný z naľožení wel. Seniorátu gemerského | Slovenská národná knižnica | Digitálna knižnica (snk.sk).

Uvedené príklady svedčia o tom, že ani problém rozpoznávania historických tlačí nástrojmi OCR stále nemožno považovať za uspokojivo vyriešený.

METÓDA HTR

Metóda HTR sa vyvíja od začiatku 21. storočia. Predmetom výskumu a vývoja je snaha čo najefektívnejšie rozpoznávať texty rukopisných dokumentov. Vzhľadom na nekonečnú varietu ľudskou rukou písaných dokumentov a množstvo písiem a rukopisných štýlov sa zdalo, že rozpoznávanie rukopisov predstavuje neriešiteľný problém. Na riešenie tohto problému sa využívajú *umelé neurónové siete* (ANN), čiže výpočtové modely využívané v oblasti umelej inteligencie, zostavené na základe abstrakcie vlastností biologických nervových systémov. Neurónové siete sa považujú za jeden z najlepších algoritmov *strojového učenia*. V rozpoznávaní rukopisných textov fungujú tak, že sa vytvorí *základný model*, ktorý sa potom cvičí na čo najväčšom množstve dát. Na základe cvičných dát sa model zdokonaľuje a rozpoznáva nové texty stále s lepšou presnosťou.

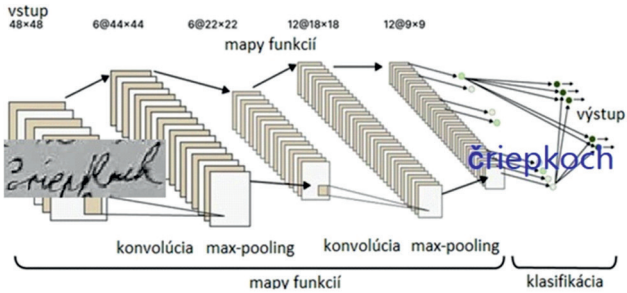
Softvér HTR+ platformy *Transkribus* zatiaľ nemôže okamžite spustiť spoľahlivý automatický prepis, ale najprv musí byť vyškolený na konkrétny typ písma a rukopisu. HTR+ je nástroj na rozpoznávanie rukopisného textu vyvinutý tímom *CITlab* na Univerzite v Rostocku. Transkripčný mechanizmus je založený na *TensorFlow*. *HTR+* je okrem stroja *PyLaia* jedným z dvoch nástrojov transkripcie, ktoré sú v súčasnosti dostupné v *Transkribe*.

Jedným z najbežnejšie používaných typov sietí na rozpoznávanie obrazu je *konvolučná neurónová sieť* alebo convnet.

OD OBRÁZKOV K PREKLADU

V snahách sprístupniť historické písomné dedičstvo sa koncentruje pozornosť výskumníkov na *transkripciu* a strojové učenie s použitím *konvolučných neurónových sietí*. Ide o proces, v ktorom sa nasnímaný *obrázok* mení na *text*. Ide o *komunikačný akt* a textotvornú operáciu – transkripciu, v ktorej vzniká *text*.

Tento text môže byť východiskom pre ďalšiu *metatextotvornú operáciu*, ktorej výsledkom je *preklad*, ktorý sa považuje za *komunikačný akt*, v ktorom je podľa Popoviča[2] preklad textotvorná operácia, ako „štylistické modelovanie prototextu jeho prekladovým metatextom“. Pravda, cesta k automatickému prekladu historických textov je rovnako zložitá ako transkripcia historických dokumentov, pretože čo najpresnejšia transkripcia si vyžaduje množstvo cvičných súborov dát. Historické novoveké texty majú svoje jazykové a štylistické špecifiká, ktoré sú pre bežné strojové prekladače ťažko zvládnuteľné.



Obrázok 5 Príklad štruktúry konvolučnej siete.

2 POPOVIČ, Anton: *Originál – preklad. Interpretačná terminológia*. Bratislava : Tatran, 1983, s. 164.

Grafika upravená podľa: Felix Dietrich. OCR vs. HTR or "What is AI, actually? OCR vs. HTR or "What is AI, actually?" - READ-COOP (readcoop.eu)

Vysvetlivka: Stroj rozpozná a transkribuje obrázok rukopisného písma na text „čriepkoch". Obrázok ilustruje vysvetlenie procesu fungovania konvolučnej siete podľa F. Dietricha[3]. Na vstupe (Input) procesu rozpoznávania nejakého predmetu, napríklad písma, tváre, zvieraťa, auta sú pixely obrázka. Vstupný obraz má rozlíšenie 48 x 48. Z neho sa vyberajú pixely. Potom sa postupne použijú množiny filtrov (Mapy funkcií – Feature Maps) na extrahovanie lokálnych obrazových príznakov prostredníctvom operácie konvolúcia (convolution), čo je matematická operácia. Filtre sú v podstate masky, ktoré sú „prehodené" cez obrázok, aby sa zistilo, či im niečo vyhovuje. Pri klasickom priraďovaní vzorov, napríklad v systémoch OCR, by človek musel vopred určiť, ako tieto filtre vyzerajú, ale v konvolučnej sieti začínajú náhodne a potom sa počas učenia zdokonaľujú. Konečný súbor funkcií sa potom vloží do husto pripojenej siete, odkiaľ pochádza skutočná univerzálna predikčná sila tohto algoritmu (classification). Takáto sieť sa môže naučiť aproximovať akúkoľvek primerane dobre vycvičenú funkciu s ľubovoľnou presnosťou, pokiaľ je sieť dostatočne veľká. V prípade transkripcie rukopisov to prakticky znamená, že na cvičenie modelu je potrebný veľký súbor cvičných dát. Otázka je, aký veľký by ten súbor mal byť, aby výsledky transkripcie boli čo najpresnejšie. V konvolučnej sieti je táto vrstva potom pripojená k sekvencii výstupov, čo môže byť naozaj čokoľvek. Pre HTR by sme v ideálnom prípade chceli mať v tejto konečnej vrstve znaky alebo dokonca slová. Počas cvičenia sa do siete vloží súbor referenčných obrázkov so známym obsahom a potom sa jeho výstup porovná so skutočnými hodnotami. Na základe rozdielu medzi predpoveďou modelu a „ground truth" sa parametre vo vnútri siete aktualizujú iteratívne. Po dokončení cvičenia je možné rozpoznať nové obrázky pri pohľade na výstup, ktorý ukazuje najpresnejšiu aktiváciu.

PRACOVNÝ POSTUP TRANSKRIPCIE

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

1. Príprava
2. Inštalácia *Transkribus* Expert Client
3. Snímanie
4. Postprocesing
5. Import dokumentov
6. Segmentácia a manuálna transkripcia
7. Tvorba modelu transkripcie
8. Sprístupnenie a použitie výsledkov transkripcie

Výskum a vývoj v oblasti rozpoznávania rukopisov už prebieha asi 20 rokov. Výsledkom tohto úsilia je množstvo poznatkov, riešení a nástrojov, s ktorými sa v projekte *SKRIPTOR* postupne zoznamujeme. Vzhľadom na relatívne malé riešiteľské kapacity a úväzky na projekte sa snažíme zvládnuť jadro vedomostí a zručností na používateľskej úrovni. Dôležité je, že riešitelia získavajú cenné *know-how* tvorby vlastných modelov transkripcie, niekedy metódou pokus – omyl. Pritom máme na pamäti hlavný cieľ

3 DIETRICH, Felix: OCR vs. HTR or "What is AI, actually?" In: *READ-COOP* [online]. Innsburck : READ-COOP SCE, 2021 [cit. 2022-11-20]. Dostupné na: https://readcoop.eu/insights/ocr-vs-htr/

a zmysel nášho úsilia. Tým je tvorba niekoľkých agregovaných, základných modelov na transkripciu historických dokumentov písaných v západoslovanských jazykoch, ako aj v iných jazykoch, ktoré sa historicky bežne vyskytovali v dokumentoch v našej kultúrnej oblasti, medzi ktoré patrí latinčina, nemčina, maďarčina. Rovnako sa usilujeme o šírenie našich skúseností vo vzdelávaní a v komunite vedcov a odborníkov z archívov, knižníc a z akademickej sféry.

PRÍPRAVA

V procese prípravy ide najmä o informačný prieskum (heuristiku) a identifikáciu a výber zbierok a dokumentov v archívoch a knižniciach. V prípravnej fáze je ďalej potrebné: a) riešenie podmienok dostupnosti zbierok a dokumentov, b) kvantifikácia a výber dokumentov na transkripciu, c) posúdenie fyzického stavu, d) zistenie stavu spracovania fondov, zbierok a dokumentov a využitia popisov vo výskume, e) dostupnosť metadát, f) zistenie fyzických rozmerov dokumentov, g) zistenie počtu strán, h) preskúmanie homogénnosti dokumentov, i) dohoda s vlastníkom alebo správcom zbierky o mieste a spôsobe snímania a o právach sprístupnenia. Proces prípravy je vhodné dokumentovať v písomnej forme, čo môže byť užitočné pri editovaní výsledkov práce.

KONTEXT

V procese prípravy je vhodné zoznámiť sa zameraním a cieľmi výskumu *SKRIPTOR*, čo je projekt APVV-19-0456 SKRIPTOR (2020 – 2024) s názvom *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*]. Riešiteľskou organizáciou je 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.). V roku 2017 sme pracovali s verziou *Transkribus Expert Client* v.1.3.7. V novembri roku 2022 bola k dispozícii verzia 1.22.1. Ako inšpirácia pre ďalších odborníkov, študentov, archivárov a iných záujemcov o transkripciu je v repozitári Slezskej univerzity v Opave k dispozícii na štúdium aj projekt HITEXT, ktorý pripravila v r. 2022 Slezská univerzita ako návrh projektu aplikovaného výskumu v programe NAKI III.

V procese prípravy je rovnako vhodné zoznámiť sa s metodologickým a technologickým kontextom transkripcie a s projektom *READ Recognition and Enrichment of Archival Documents*. Ide o projekt, ktorého riešenie prebiehalo v rokoch 2016 – 2019 v rámci programu Horizon 2020.[4] Výskum bol predtým financovaný ako súčasť projektu *tranScriptorium.* Tento projekt získal finančné prostriedky zo siedmeho rámcového programu Európskej únie pre výskum, technologický rozvoj podľa dohody o grante č. 600707. Rovnako je užitočné poznať združenie *READ-COOP*, ktoré v októbri roku 2022 malo 113 členov z 27 krajín a v novembri 2020 už 100 000 používateľov. Taktiež je dôležité zoznámiť sa s fungovaním kreditového systému a so spôsobom používania

4 READ Recognition and Enrichment of Archival Documents. In: *CORDIS: EU research results* [online]. Last update 17 August 2022 [cit. 2022-11-20]. Dostupné na: https://cordis.europa.eu/project/id/674943

kreditov na transkripciu. V *READ-COOP* sa kupujú kredity. Nejde o zisk združenia, ale o príjem, ktorý sa používa na výskum, vývoj a infraštruktúru. V projekte *SKRIPTOR* „spotrebujeme“ na transkripciu asi 10 000 kreditov, čo znamená asi 2000 €. Každý nový používateľ má automaticky k dispozícii na samotnú transkripciu bezplatných 500 kreditov. Na nahrávanie, segmentáciu, cvičenie modelu sa kredity nemíňajú. Používanie kreditov treba starostlivo zvažovať pri cvičení modelu. Na cvičenie modelu by sa mali používať len menšie vzorky strán dokumentu. Masovú transkripciu je najlepšie spustiť až vtedy, keď sme relatívne spokojní s vytvoreným modelom transkripcie.

FONDY A ZBIERKY

Z vecného hľadiska sú pre projekty transkripcie dôležité poznatky o archívnych fondoch a zbierkach. Historické rukopisné**,** prípadne strojopisné dokumenty na transkripciu sa nachádzajú prevažne v archívoch. Historické tlačené dokumenty sa nachádzajú hlavne v knižniciach, ale aj u iných právnických alebo fyzických osôb. Na usporiadanie archívnych fondov sa u nás používa *Klasifikačná schéma archívnych fondov a zbierok štátnych archívov na Slovensku*. Na najvyššej úrovni majú archívy spravidla svoje *zoznamy archívnych fondov a zbierok*. Tieto zoznamy obsahujú všeobecné atribúty fondu a zbierky: názov archívneho fondu/zbierky, časové rozpätie, rozsah veľkosti archívneho fondu/zbierky v bežných metroch, prístupnosť a typ archívnej pomôcky. Výber konkrétnych dokumentov na transkripciu a výskum záleží na erudícii výskumníka, pretože rozsah a hĺbka spracovania fondov a zbierok je rôzna.

Podmienkou transkripcie archívnych dokumentov je digitalizácia. Na Slovensku sa komplexnejšie venoval digitalizácii archívov R. G. Maretta.[5]

Archívne fondy a zbierky často predstavujú obrovské množstvá dokumentov. Na Slovensku štátne archívy uchovávajú asi 200 kilometrov dokumentov. Podľa výskumu Medzinárodnej archívnej asociácie (ICA) z roku 2020[6] však 3 % archívov nemajú spracované zbierky vôbec a 50 % archívov má zbierky spracované na základnej úrovni. Výskum tiež ukázal nízku mieru využívania medzinárodných štandardov odporúčaných ICA. Ide o štandardy spracovania: ICA ISAD(G), ISDF, ISAAR (CPF) a ISDIAH. Najrozšírenejší je štandard ISAD(G).

ŠTANDARD ISAD (G)

Medzinárodný štandard ISAD(G)[7] definuje zoznam prvkov a pravidiel na popis archívov a popisuje druhy informácií, ktoré musia a mali by byť zahrnuté v takýchto opisoch. Vytvára hierarchiu popisu, ktorá určuje, aké informácie by mali byť zahrnuté

5 MARETTA, Gregor Robert: Digitalizácia stredovekých listín v Slovenskom národnom archíve. In: *Slovenská archivistika*, roč. 34, č. 1, 2009, s. 16 – 40.

6 International Council on Archives: Archival Arrangement & Description : Global Practices. Report on the survey undertaken by the ICA Training Programme, with a foreword by the ICA President [online]. July 2020 [cit. 2023-01-20]. Dostupné na: https://www.ica.org/sites/default/files/aad_survey_report_final_202108_eng.pdf

7 *ISAD(G) : general International Standard Archival Description : adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 September 1999* [online]. Ottawa, 2000 [cit. 2022-11-20]. Dostupné na: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

na akej úrovni. V súvislosti s výskumom a experimentami s transkripciou archívnych dokumentov považujeme za vhodné, aby boli transkribované fondy, zbierky a dokumenty popísané na štandardnej úrovni. Tento štandard poskytuje rámec pre spoločný prístup a nie rigidný formát. Dostupný je český preklad.[8] Existuje aj neoficiálny slovenský preklad.[9]

Princípy popisu fondov, zbierok a dokumentov v ISAD(G) sa riadia štyrmi všeobecnými zásadami: *1) Opis od všeobecného po konkrétny* – Viacúrovňový popis sa začína od všeobecnej úrovne popisu, ktorá je zvyčajne fondmi, a pokračuje do podrobnejších úrovní, ako sú podfondy, séria, súbor, položka atď. Táto hierarchická štruktúra musí byť reprezentovaná a správne definovaná v archívnom opise. *2) Informácie relevantné pre úroveň opisu* – Informácie na každej úrovni opisu sa musia týkať len archívnej jednotky opísanej na tejto úrovni. *3) Prepojenie popisov* – Každá archívna jednotka musí byť prepojená so svojou nadradenou úrovňou v rámci hierarchie a jej úroveň musí byť explicitná. *4) Neopakovanie informácií* – Aby sa zabránilo opakovaniu musia sa všeobecné informácie spoločné pre skupinu deklarovať na najvyššej možnej úrovni. Podúrovne musia zase obsahovať spoločné informácie, ktoré sa vzťahujú na jej nižšie úrovne.

ISAD(G) definuje dvadsaťšesť dátových údajov popisu archívnych fondov, zbierok a dokumentov. Vo výskume *SKRIPTOR* sa odporúča použiť na popis zbierok a dokumentov tieto pravidlá s návestiami a v ďalej uvedenom poradí. Popisy je možné tiež použiť pri editovaní zbierok, špecifikácii metadát v *Transkribus Expert Client* a na prezentáciu transkribovaných zbierok v nástroji *read&search*.

V prvej oblasti **1. Vyhlásenie o totožnosti** týchto 6 povinných údajov identifikujúcich fond, zbierku, dokument: 2. Referenčné kódy: Prvky používané na jednoznačnú identifikáciu jednotky opisu: kód krajiny, kód úložiska, špecifický miestny referenčný kód/kontrolné číslo/iný jedinečný identifikátor. 3. Titul: Názov jednotky opisu. 4. Dátum: Dátumy vytvorenia záznamu počas vedenia záležitostí alebo dátumy vytvorenia dokumentu. 5. Úroveň popisu: Úroveň jednotky opisu v rámci hierarchie. 6. Rozsah a médium jednotky opisu: Fyzikálny alebo logický rozsah a médium jednotky opisu. Ďalej sú oblasti popisu.

V oblasti **2. Kontext** je povinný len údaj Meno: Tvorca jednotky popisu. Ďalšie údaje v oblasti kontext sú: Administratívne/biografické dejiny: Biografické alebo administratívne podrobnosti týkajúce sa tvorcov jednotky opisu; Archívna história: Príslušné historické informácie o jednotke opisu; Bezprostredný zdroj akvizície alebo prevodu: Zdroj získania materiálu.

V oblasti **3. Obsah a štruktúra** sú údaje: **Rozsah pôsobnosti a obsah:** Zhrnutie rozsahu a obsahu relevantného pre úroveň opisu; **Informácie o hodnotení, zničení a plánovaní:** Zhodnotenie, zničenie a plánovanie činností vykonaných alebo plánovaných pre jednotku opisu. **Časové rozlíšenie:** Plánované dodatky k jednotke opisu.

V oblasti **4. Podmienky prístupu a používania** sú údaje**: Podmienky upravujúce prístup:** Informácie o právnom postavení, ktoré môžu ovplyvniť prístup k jednotke opisu; **Podmienky, ktorými sa riadi reprodukcia:** Podmienky pre reprodukciu jednotky

8 *ISAD(G) : Všeobecný mezinárodní standard pro archivní popis. Přijato Komisí pro popisné standardy, Stockholm, Švédsko, 19. – 22. září 1999* [online]. Praha, 2009, 57 s. [cit. 2022-11-20]. Dostupné na: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

9 DRAŠKABA, Peter – HANUS, Jozef: Všeobecná medzinárodná norma pre opis archívnej jednotky. In: *Slovenská archivistika*, roč. 35, č. 1, 2000, s. 197 – 215.

opisu po vytvorení; **Jazyk/skripty materiálu: Jazyky, skriptá** a systémy symbolov používané v jednotke opisu; **Fyzikálne vlastnosti a technické požiadavky:** Príslušné fyzické podmienky, softvérové a hardvérové požiadavky na prístup k jednotke opisu a jej uchovávanie; **Hľadanie pomôcok:** Nájdenie pomôcok použiteľných pre jednotku opisu.

V oblasti **5. Spojené materiály** sú údaje**: Existencia a umiestnenie originálov:** Informácie o existencii alebo zničení pôvodnej jednotky opisu; **Existencia a umiestnenie kópií:** Informácie o existencii a dostupnosti kópií jednotky opisu; **Súvisiace jednotky opisu:** Informácie o jednotkách opisu súvisiacich s pôvodom alebo inými asociáciami s jednotkou opisu; **Poznámka k uverejneniu:** Publikácie, ktoré sa týkajú alebo sú založené na použití, štúdiu alebo analýze jednotky opisu.

V oblasti **6. Poznámky** sú informácie, ktoré sa nezmestia do žiadnej z predchádzajúcich oblastí.

V oblasti **7. Kontrola** sú údaje: **Poznámka archivára:** Informácie o tom, kto a ako pripravil popis. **Pravidlá alebo dohovory:** Protokoly, na ktorých je opis založený. **Dátum(-y) popisu:** Dátumy vytvorenia a revízie.

INŠTALÁCIA TRANSKRIBUS EXPERT CLIENT

Z metodologického hľadiska je pre výskumy a experimenty s transkripciou v projekte *SKRIPTOR* najvýznamnejšia platforma *Transkribus.* Návody na prácu s platformou *Transkribus sú podrobne spracované v dokumentácii.*[10]

Transkribus je 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. 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 roku 2022 mal *Transkribus* viac ako 100 000 používateľov, 40 mil. obrazov, 20 mil. rozpoznaných strán. Platforma bola vytvorená v kontexte dvoch projektov EÚ *tranScriptorium* (2013 – 2015) a *READ* (2016 – 2019).

ALTERNATÍVY *TRANSKRIBUS*

V projekte *SKRIPTOR* 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 *eScriptorium*, 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*.[11] 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*

10 Resource center. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-11-20]. Dostupné na: https://readcoop.eu/transkribus/resources/

11 *Projekt Pero* [online]. [cit. 2022-11-20]. Dostupné na: https://pero.fit.vutbr.cz

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.

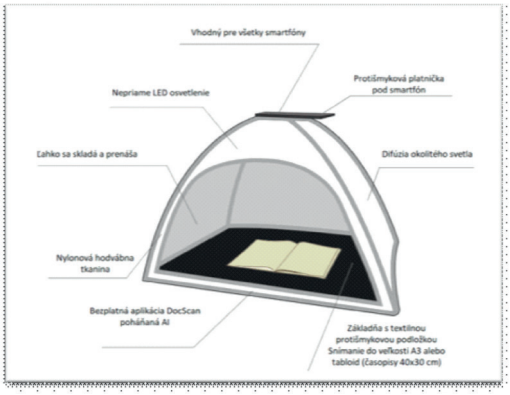
SNÍMANIE

Snímanie je jeden z procesov *digitalizácie*. Zahŕňa skenovanie, fotografovanie dokumentov, pomenovanie a organizáciu adresárov a súborov v počítači, archivovanie zdrojových súborov a zálohovanie derivovaných súborov.

Snímanie sa vykonáva pomocou vhodného technického zariadenia na digitalizáciu. Ide o zariadenie na zachytenie *digitálneho obrazu* (digitálne fotoaparáty a kamery, skenery na knihy, dokumenty alebo mikrofilmy, audio- a videohardvér) pripojené na vhodnú počítačovú platformu (počítač, operačný systém, sieť). Rozlišujeme dve rôzne metódy snímania: a) *skenovanie* a b) *fotografovanie*, čiže používanie *digitálnych kamier/ fotoaparátov*, *mobilných telefónov*. V súlade s predpisom Ministerstva vnútra SR[12] archív *„d) umožňuje snímanie archívnych dokumentov a priestorov archívu klasickou kamerou a digitálnou kamerou (ďalej len „kamera") a fotografickou technikou*". Na účely automatickej transkripcie, pokiaľ je to možné, použijeme dokumenty nasnímané profesionálnymi skenermi a obrazmi v najvyššej dosiahnuteľnej kvalite. Minimálna kvalita skenovania by mala byť 300 DPI. Pri historických rukopisoch ide *de facto* o grafiku, preto je vhodné skenovať vo vyššej kvalite.

SCANTENT A DOCSCAN

Pri platforme *Transkribus* je možné snímať dokumenty do fomátu veľkosti A3 zariadením *ScanTent* so softvérom *DocScan*.



Obrázok 6 Zariadenie na snímanie fotografovaním dokumentov pomocou smartfónov.

12 Oznámenie generálneho riaditeľa sekcie verejnej správy o uverejnení opatrenia ministra vnútra Slovenskej republiky č. SVS-204-2008/00111 zo 17. januára 2009 o správnych informáciách a službách štátnych archívov zriadených Ministerstvom vnútra Slovenskej republiky, čl. 10. In: *Vestník Ministerstva vnútra Slovenskej republiky*, roč. 2009, čiastka 6, 29. január 2009.

Obrázok 7 Popis dokumentu v DocScan. Nasnímaním QR kódu sa popíše dokument, ktorý snímame.

DocScan je open source aplikácia pre Android navrhnutá pre *ScanTent.* Identifikuje strany dokumentu v živom náhľade a robí snímky v dostatočnej kvalite na transkripciu. V automatickom režime nasníma obrázok po otočení stránky. Umožňuje rýchlo skenovať knihy alebo dokumenty bez interakcie s mobilom. Obrazovku smartfónu je možné zdieľať na obrazovke počítača a vzdialene ovládať smartfón napríklad cez *TeamViever*. Vďaka spoločnosti *Ifunplay* a aplikácie *DocScan* možno teraz *ScanTent* používať aj s operačným systémom *iOS v iPhonoch*. Držiak na vrchu *ScanTent* umožňuje umiestnenie smartfónu a optimálny pozorovací uhol a konštantnú vzdialenosť. Ak denné svetlo nestačí, biele LED pásiky poskytujú rovnomerné osvetlenie, ktoré maximalizuje kvalitu obrazu. Používanie týchto jednoduchých praktických zariadení sa postupne rozširuje. Francúzska národná knižnica používa 40 zariadení *ScanTent*.

V projekte *SKRIPTOR* sa stretávame s dokumentmi, ktoré už boli nasnímané v archívoch alebo v knižniciach a podľa našich skúseností sú väčšinou vhodné na transkripciu. Tieto digitálne súbory však majú rozličnú kvalitu a rôzne formáty. Často je však potrebné dokumenty snímať priamo v inštitúciách. V mnohých archívoch a knižniciach, bádateľniach a študovniach nie sú k dispozícii žiadne snímacie zariadenia a archívne dokumenty nie je možné vynášať z archívov. Používatelia niekedy snímajú dokumenty alebo ich časti len podľa svojich ohraničených možností a v malom rozsahu na svoje výskumné ciele „z ruky“, mobilnými telefónmi a fotoaparátmi bez akýchkoľvek pomôcok a pri premenlivom osvetlení. Východiskom pohodlnejšieho snímania fotografovaním by mohli byť v bádateľniach archívov a v študovniach knižníc systémy *ScanTent a DocScan*.

FORMÁTY OBRÁZKOV

Fotografie je možné tvoriť, ukladať a upravovať v rôznych *formátoch*.[13] Najčastejšie ide o súbory vo formátoch *RAW a JPG*. Z hľadiska úprav fotografií je dôležitý formát *TIFF*. Najrozšírenejší je *formát JPG či JPEG*, ktorý sa vyskytuje s príponou *.jpg, .jpeg* alebo *.JPG, JPEG.* Medzi nimi nie je žiadny rozdiel. V tomto formáte ukladajú súbory všetky

13 LIŠKA, Matej: *Zoner Photo Studio X. Praktická príručka*. Brno : Zoner software, 2021. 178 s.

fotoaparáty. V niektorých je možné voliť jeden formát alebo snímanie v dvoch formátoch *JPG a RAW, ARW.* Výhodou formátu *JPG* je, že sa obrázok dá zobraziť prakticky v každom zariadení – v mobilnom telefóne, televízore alebo vo webovom prehliadači. Zaberá málo miesta na disku, je úsporný, pretože ide o kompresiu so stratou. Nevýhodou tohto formátu je, že každou úpravou obrázok stráca kvalitu pri každom uložení. V projektoch transkripcie používame na snímanie mobilnými zariadeniami formát *JPG* na archivovanie a v transkripcii spravidla pracujeme s derivovaným formátom *PDF*.

Formát RAW znamená, že nasnímaný súbor je „surový", nespracovaný a dáta nie sú komprimované. Dáta v tomto formáte sú veľmi veľké a na ich spracovanie je potrebný špeciálny softvér, napríklad komerčný *Zoner Photo Studio* alebo open source *FastStone Image Viewer*. Výsledné obrázky majú vysokú kvalitu a sú po úprave vhodné na kvalitné editovanie.

Formát TIFF sa vyskytuje s príponami *.tiff, tif.* Pri ukladaní do tohto formátu spravidla nedochádza ku kompresii dát. Ak áno, tak ide o bezstratovú kompresiu aj pri opakovanom ukladaní. Súbor zachováva maximum informácií z formátu *RAW* pri editácii. Nevýhodou je veľkosť súborov vo formátoch *TIFF.* V profesionálnych projektoch digitalizácie je formát *TIFF* najvhodnejší na dlhodobé archivovanie.

V procesoch snímania je nevyhnutné zvoliť metódu *zálohovania a archivovania* zdrojových obrázkov a ich derivátov. Základné pravidlo o zálohovaní vyžaduje urobiť najmenej tri kópie na dva rôzne *nosiče* a jednu – archívnu zálohu mať na vzdialenom mieste. Každá snímka by mala mať aspoň dve kópie, a to na dvoch rôznych úložiskách, napríklad na SD karte, disku, externom disku, digitálnom repozitári.

POSTPROCESING

Obrázky *získané vlastným snímaním alebo z iných zdrojov je pred nahrávaním na server Transkribus* potrebné *následne upraviť*. Ide o „*postprocesing*". V následnom spracovaní prezrieme nasnímané obrázky v súboroch, odstránime duplicity, zisťujeme, či je potrebné niektoré strany opätovne snímať, upravíme orientáciu strán, skontrolujeme kvalitu a komplexnosť dokumentu, pridelíme zbierke a dokumentom názvy, vytvoríme zálohy súborov, vytvoríme deriváty formátov, prehľadne uložíme deriváty dokumentov vo vhodnom formáte do adresárov na svojom počítači tak, aby sme ich mali pripravené na nahrávanie (loading) na server.

Vo výskume *SKRIPTOR* pracujeme so systémom *Transkribus Expert Client*. Po vytvorení vlastného používateľského účtu nasleduje inštalácia *Transkribu* na osobný počítač, vytvorenie vlastnej zbierky a nahrávanie zvolených súborov do vlastnej zbierky.

IMPORT DOKUMENTOV (UPLOAD)

V štruktúre systému *Transkribus Expert Client* sú dva kľúčové prvky: *zbierky a dokumenty*. *Zbierka* je nadradená *dokumentu*. Dokumenty sú usporiadané do tzv. zbierok. Zbierky možno chápať ako priečinky obsahujúce dokumenty. Zbierky sa zvyčajne tvoria podľa nejakého konkrétneho projektu. Napríklad všetky dokumenty patriace k jednému projektu sú usporiadané do jednej zbierky. Napr. zbierky: Koháry korešpondencia, Collectanea Laučeka, Hurban listy, Abrahamides kázne, Reambulačné protokoly, Rukopisné katalógy, Andrej Kmeť

korešpondencia a pod. Jedna zbierka môže obsahovať viac dokumentov, ktoré pozostávajú z jednej alebo viacerých strán. Každá zbierka v *Transkribe* má jedinečný *identifikátor* (ID). Každý dokument v zbierke má jedinečný číselný identifikátor, názov dokumentu, počet strán dokumentu, meno osoby, ktorá nahrala dokument do *Transkribu*, dátum a čas nahratia, vlastníka zbierky. Zbierku je možné manažovať – tvoriť, vymazať, upravovať, pridávať a upravovať oprávnenie používateľov zbierky so súhlasom a rozhodnutím vlastníka zbierky, pracovať s kreditmi k zbierke. Ku každému dokumentu je možné popísať všeobecné metadáta a metadáta ku jednotlivej strane, ako aj štrukturálne a textové metadáta a komentáre. Používateľ môže mať niekoľko zbierok s rôznymi dokumentmi. Na účely prezentačnej vrstvy nazvanej *read&search* je potrebné vytvoriť *jednu spoločnú zbierku* s viacerými dokumentmi. Všetky zbierky a dokumenty v *Transkribe* sú súkromné.

Po vytvorení zbierky je potrebné v *Transkribe* nahrať *dokumenty*. Potom je možné spustiť nástroje, ako je analýza rozloženia (segmentácia) alebo rozpoznávanie textu (transkripcia). Údaje v *Transkribe* sú vždy súkromné a prístupné iba jednotlivým používateľom. *Vlastník* zbierky (owner) môže umožniť prácu aj iným používateľom (users) s oprávneniami, ktoré im pridelí (owner, editor, transcriber, reader).

Na server sa nahrávajú *dokumenty*. V štruktúre systému *Transkribus Expert Client* je dokument zaradený do nejakej zbierky. Dokument môže byť presunutý do inej existujúcej zbierky. Základné metadáta k dokumentu sú: jedinečný číselný identifikátor, názov dokumentu, meno osoby, ktorá nahrala dokument do zbierky v *Transkribe*, dátum a čas nahratia do zbierky, meno zbierky, do ktorej dokument patrí. Dokument je možné zobraziť vo forme „Overview“ s jednotlivými stranami a grafickým rozlíšením stavu stránky (napr. Ground Truth, In progress, Done, Final). Po nahratí dokumentu má každá strana dokumentu status „new“. Vo forme „Layout“ sú viditeľné texty transkripcie strán, riadky textu, poradie čítania riadkov strojom, identifikátor riadka a koordináty umiestnenia elementov v riadku.

TRANSKRIPCIA ČI TRANSLITERÁCIA?

V súvislosti s metódami rozpoznávania znakov v historických textoch sa často zamieňajú dva termíny: *transkripcia a transliterácia*.

Termín *transliterácia* (odborný/vedecký prepis, odborná/vedecká transkripcia, zriedkavo prepísmenkovanie) sa v jazykovede definuje a) po písmenách uskutočňované „pretlmočenie” textov či slov zapísaných jedným grafickým systémom prostredníctvom iného grafického systému; b) grafický prepis cudzojazyčného textu nahradzujúci písmená jednej abecedy písmenami druhej abecedy bez prihliadnutia k fonetickej hodnote, takže je možný spätný prepis; c) prevod z jedného grafického systému do druhého, pričom každému písmenu jedného grafického systému zodpovedá vždy písmeno druhého systému (rovnaké písmeno alebo spojenie písmen), takže je možný aj jednoduchý spätný prevod do jazyka originálu. Podobná je definícia: Prepis z jedného písma do druhého písma iného jazyka, pri ktorom sa zachováva jednoznačnosť zápisu medzi písmenami obidvoch abecied. Transliteráciou sa podľa Pravidiel slovenského pravopisu (2013) rozumie napríklad prepis zo slovanských jazykov používajúcich *cyrilské písmo do písma slovenčiny* (do latinského písma používaného v slovenčine). Teda *prepis znakov, písiem z jedného jazyka (napríklad ruštiny) do iného jazyka (slovenčiny) alebo prepis z iných grafických sústav.* Pravidlá

transliterácie cyrilských písmen zo súčasných slovanských jazykov, ktoré používajú cyriliku, určujú slovenské technické normy. Prepisy z najdôležitejších jazykov Ďalekého východu, a to z čínštiny, japončiny a kórejčiny upravujú prepisy, v ktorých sa čo najvernejšie zachytáva ich zvuková podoba. V platforme *Transkribus* sa viac konvenčne ako presne používa termín transkripcia. Termíny transliterácia a transkripcia sa používajú často v rovnakom význame.

Transkripcia (prepis) môže byť podľa Wikipédie: a) v užšom zmysle: písomné vyjadrenie (vyslovovaných alebo cudzím grafickým systémom napísaných) slov a textov z hľadiska ich výslovnosti prostriedkami určitého grafického systému (v najužšom zmysle len takéto písomné vyjadrenie slov a textov napísaných cudzím grafickým systémom); b) v širšom zmysle: bod a) plus transliterácia; c) v najširšom zmysle: vyjadrenie výrazu jedného jazyka v inom grafickom systéme ako v grafickom systéme, v ktorom sa tento jazyk obyčajne zapisuje, resp. v ktorom je už zapísaný (táto definícia zahŕňa okrem bodu B napr. prevod slovenského textu do Braillovho písma, prevod posunkovej reči do písanej notácie a [pravdepodobne aj] moderný prepis textu archiválie).

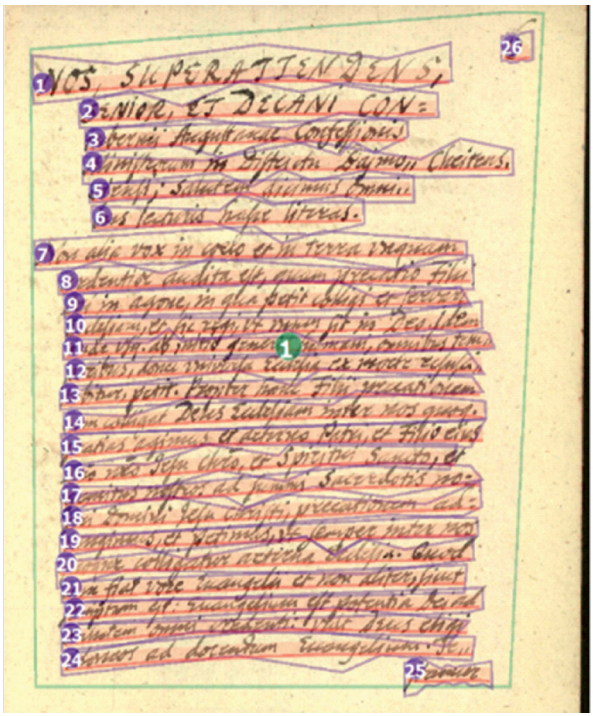
V platforme *Transkribus* sa používa termín transkripcia vo význame prepisu rukopisného alebo tlačeného historického textu *v určitom jednom jazyku* a automatický prepis textu *v tom istom jazyku*. Napríklad rukopis v maďarčine sa prepisuje pomocou znakového súboru tlačenej latinky. Nejde teda o prepis ani preklad medzi jazykmi, ale o prepis v rámci jedného jazyka.

SEGMENTÁCIA A MANUÁLNA TRANSKRIPCIA

Po nahratí obrázkov na server nasleduje *analýza rozloženia* (Layout Analysis). Jednotlivé nahraté dokumenty v zbierke majú v nástroji *Transkribus Expert Client* formu obrázkov, ktoré vznikli v procese snímania (skenovania). Sú to obrázky stránok dokumentov nahratých do platformy *Transkribus* napríklad vo formáte *PDF, JPG, PNG, TIFF*. Obrázky je potrebné *segmentovať*, identifikovať jednotlivé prvky obrázkov. Na účely transkripcie dokumentu je najprv potrebné obrázok rozdeliť na *textové oblasti a riadky* (TR – Text Regions a Lines). Ide o uplatnenie metódy analýzy obrazu a textovej analýzy, pričom výsledkom tejto analýzy je členenie stránky textu na časti stránky – analýzou sa vyznačujú hlavne bloky textu, horizontálne členenie textu, podstatné, prípadne okrajové, nadbytočné časti obrazu, riadky a základné čiary. Analýzu rozloženia je možné vykonať niekoľkými kliknutiami a vo väčšine prípadov si úkon nevyžaduje manuálne opravy. To závisí od zložitosti štruktúry vstupného dokumentu. V *Transkribus Lite* sa analýza rozloženia (segmentácia) spustí automaticky, keď sa spustí úloha rozpoznávania textu. Automatická pokročilá analýza rozloženia *CITlab* vo svojom štandardnom nastavení zvyčajne rozpozná jednu *oblasť textu* (TR) na obrázku so zodpovedajúcimi základnými čiarami. Existujú však aj rozloženia, pri ktorých sa odporúča použitie viacerých textových oblastí. Ide o situácie, keď existujú poznámky na okraji alebo poznámky pod čiarou a podobné opakujúce sa prvky. Pokiaľ sú tieto textové oblasti, ktoré sa líšia obsahom a štruktúrou, obsiahnuté v jednej textovej oblasti, analýza rozloženia jednoducho počíta riadky zhora nadol. Toto poradie čítania nezohľadňuje, kam text skutočne patrí z hľadiska obsahu, ale len to, kde sa na stránke graficky nachádza. Oprava automaticky vygenerovaného, ale neuspokojivého poradia čítania môže byť časovo náročná. Problému možno ľahko predísť vytvorením niekoľkých textových oblastí (TR).

Ak chceme vygenerovať prepis HTR, musíme dokumenty rozdeliť na *textové oblasti, riadky a základné čiary.* V predvolenom nastavení je oblasť textu obdĺžnik, ktorý obklopuje všetok ručne písaný text obsiahnutý na obrázku. Je však možné upraviť textovú oblasť podľa všeobecného rozloženia pridaním kontrolných bodov, čím sa vytvorí polygón. Historické dokumenty majú niekedy zložité usporiadanie a pozostávajú z rôznych rozložení, čo môže viesť k problémom s poradím čítania prvkov textu. Pri komplikovaných rozloženiach si rýchlo všimneme, že ručne nakreslené textové oblasti sa môžu prekrývať. Tento problém sa dá ľahko vyriešiť úpravou pravouhlých oblastí textu, pridaním bodov a tým vytvorením *polygónov*.

V systéme *Transkribus Expert Client* poradie čítania zobrazuje na segmentovanej stránke to poradie, v ktorom bude stroj transkripcie čítať riadky textu na obrázku stránky. Toto poradie čítania sa vytvára automaticky počas analýzy rozloženia, ale možno ho neskôr zmeniť aj manuálne po zvolení zobrazenia čísiel riadkov. Pri automatickej analýze rozloženia je poradie čítania určené súradnicami riadkov na obrázku: horný riadok, ktorý je najviac vľavo, je číslo jeden atď. Dôležité je vedieť, že poradie čítania nie je relevantné pre samotné školenie, ale môže sťažovať čítanie transkribovanej strany. Ak sa má prepis exportovať a ďalej použiť na vydanie, tak poradie čítania je potrebné zadať správnym spôsobom, aby bol text a jednotlivé prvky textu v správnom poradí. Dá sa to jednoducho urobiť zapnutím poradia čítania na karte „viditeľnosť tvaru“. Všetky riadky tak zobrazujú kruh s číslom, ktoré označuje ich polohu na stránke dokumentu. Kliknutím na tieto krúžky sa otvorí okno s textovým editorom, kde je možné priradiť nové, správne čísla. Táto funkcia je užitočná najmä v dokumentoch s náročným rozložením, kde sa poradie riadkov neriadi bežnými pravidlami.



Obrázok 8 Prijateľná automatická analýza rozloženia (segmentácia) strany.

Vysvetlivka:

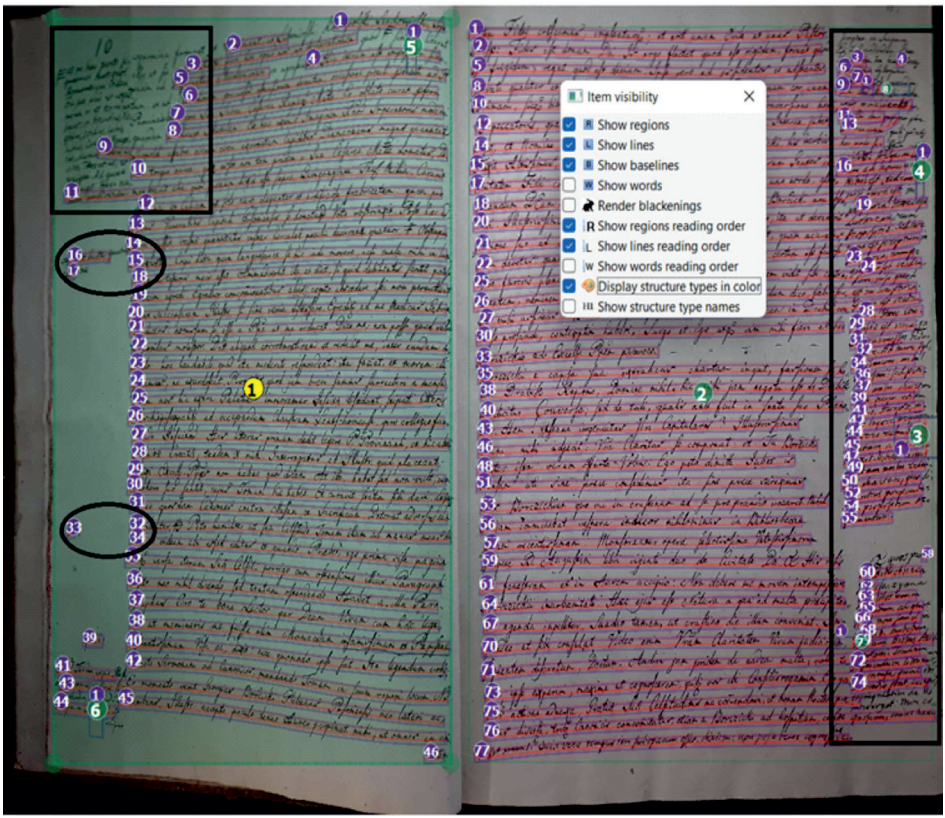
Blok textu TR je ohraničený zelenou linkou. Určené sú základné čiary a polygóny. Správne je číslovaných 25 riadkov. Manuálne by sa malo upraviť poradie riadkov (26 zmeniť na 1).

Po automatickej segmentácii strán sú potrebné dodatočné úpravy. Spustenie automatickej analýzy (segmentácie) rozloženia stránky a textu neposkytuje vždy vyhovujúce výsledky. Niekedy sú preto potrebné manuálne korekcie rozloženia. Na úpravy použijeme nástroje *editora Transkribus* – tzv. *„canvas“* a nástroje v časti editora. *Plátno „Canvas“* je menu úprav v *Transkribus* Expert Client. V ponuke Canvas, ktorá sa nachádza spravidla na ľavej strane stránky dokumentu, sa nachádzajú potrebné voľby, ako ohraničiť blok textu (TR - Text Regions), pridať riadok (L - Lines), pridať základnú linku (BL – Base Lines), pridať slovo (W – Word), pridávanie rôznych častí (tabuľky, reklamy, schémy, grafy, grafiky atd.). V ponuke Canvas je možné tiež zmeniť existujúce tvary.

Najdôležitejším referenčným bodom na rozpoznávanie textu *je Základná čiara (Baseline)*. Popisuje polyčiaru, ktorá sa tiahne pozdĺž spodnej časti rukou písaného textového riadku. Segmentáciu textu na riadky a základné čiary je možné vykonať automaticky pomocou *CITlab Advanced LA*. Pri zložitých rozloženiach a v závislosti od konkrétneho písma v rukopisoch sa však môžu vyskytnúť prípady, keď je potrebné vykonať niektoré manuálne opravy. Základná čiara by mala prebiehať pozdĺž spodnej časti textového riadku, písmená by na nej mali sedieť a zostupne smerovať nižšie. Základná čiara pozostáva z jednotlivých bodov, ktoré je možné nastaviť pri manuálnej úprave sekvencie; nastavenie sa dokončí dvojitým kliknutím alebo voľbou *Enter* v poslednom bode.

Základné línie je možné nakresliť aj vertikálne. Na obrázku a dokonca aj v textovej oblasti je možné tiež kombinovať rôzne smery čiar (napr. typické „pohľadnicové rozloženie“). Ak sa vykonávajú zmeny na linkách, je dôležité, aby sa vždy robili na základných čiarach, pretože pre každý riadok v dokumente je na pozadí aj oblasť čiary. Dajú sa zobraziť pomocou tlačidla viditeľnosti položky. Tieto riadkové regióny sa nesmú meniť, automaticky sa prispôsobia, keď niečo zmeníme na základnej úrovni. Zobrazí sa okno s otázkou, či by ste chceli zmeniť aj nadradený riadok, čo treba potvrdiť.

Riadkové oblasti (Line Regions LR) sa nachádzajú v rámci bloku textu a možno ich opísať ako mnohouholníky, v ktorých je všetok ručne napísaný text v riadku. Keďže nemajú pre proces transkripcie bezprostredný význam, riadkové oblasti by sa nemali opravovať. Ak sa niečo má zmeniť v rozložení riadkov dokumentu, vždy to treba urobiť na základnej úrovni (Baseline).



Obrázok 9 Komplikované rozloženie textu na obrázku. Problematická automatická segmentácia.

Vysvetlivky:

Bloky textu – na ľavej strane je správne zelenou linkou ohraničená oblasť s textom (TR), na pravej strane je nesprávne ako jeden blok označená časť so základným textom a vpravo obsažné marginálie na celú výšku strany. Spoločný TR textu hlavného textu a textu na margu spôsobí, že stroj číta riadky v rámci jedného bloku. Riadky 1, 2 sú v hlavnom texte a čítanie by pokračovalo číslami riadkov 3, 4 na margu. Margo je potrebné manuálne označiť ako samostatný blok textu (TR) so samostatným číslovaním riadkov.

Všetky riadky majú automaticky určenú základnú čiaru.
Čísla blokov textu sú v jednom žltom a v ďalších zelených krúžkoch. V segmentácii automat identifikoval spolu 7 blokov. Vyžaduje sa manuálna úprava.
V ľavom hornom štvorci je viditeľná neostrá časť textu, ktorá do segmentácie nebola zahrnutá.
V hornej časti ľavej strany vidno nesprávne číslované riadky. Nesprávne číslované poradie riadkov je aj v oválach. Vyžaduje sa manuálna korekcia.
Na pravej strane je vložená informácia Item visibility, v ktorej je možné nastaviť, ktoré elementy sa majú zobraziť. Voliteľné je aj farebné odlíšenie.
Problematická segmentácia komplikuje jednak manuálnu transkripciu a po transkripcii sťažuje čítanie a porozumenie textu.

Problematická segmentácia rozloženia elementov obrazu nemá vplyv na funkciu samotného stroja

transkripcie, pretože stroj v automaticky identifikovanom poradí transkribuje naučené znaky bez ohľadu na človekom očakávanú postupnosť a význam textu.

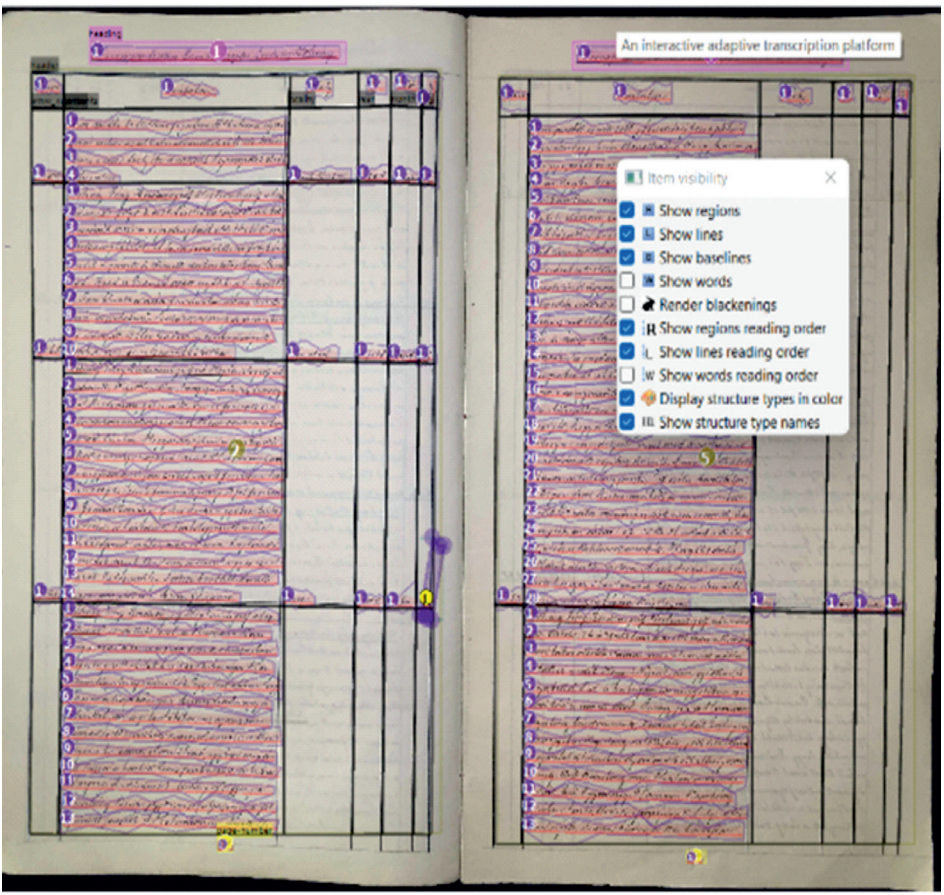
V procese úprav strán v dokumente sa menia a zaznamenávajú zmeny každej strany. Ide *o statusy transkripcie*: *New* (nový – stav pre novonahraté dokumenty), *In Progress* (prebieha – automatická zmena stavu po úprave strany), *Done* (hotovo – stránka je prepísaná), *Final* (finálna verzia – stránka prepísaná 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.

V štruktúre systému *Transkribus Expert Client* je možné pomocou funkcie štrukturálneho značkovania vo funkcionalite „metadáta" označiť, „značkovať" (mark-up) prvky štruktúry dokumentov. Ide o *tagovanie štruktúry***.** Navyše je možné cvičiť modely tak, aby automaticky rozpoznali štruktúru dokumentov. Pridaním tagov, teda štrukturálnych značiek sa vytvoria cvičné dáta pre tento proces. Nie je potrebné označovať každý prvok dokumentov – stačí sa zamerať na označenie sekcií, ktoré nás zaujímajú. Rozhranie štrukturálneho označovania v *Transkribe* umožňuje rozdeliť dokumenty do štruktúrnych sekcií, ako sú odseky, nadpisy alebo čísla strán, pridať prispôsobené kategórie značiek pre vaše individuálne potreby a v budúcnosti použiť tieto štrukturálne informácie na cvičenie modelu.

TABUĽKY

Tlačené a ručne kreslené tabuľky sú bežné v historických dokumentoch všetkých typov. V súčasnosti sa tabuľky musia v *Transkribe* kresliť ručne pomocou editora tabuliek. Technológia, ktorá umožní automatické rozpoznávanie tabuliek, je vo vývoji. V súčasnosti ide v práci s tabuľkami o poloautomatický proces. Na účely transkripcie je najprv potrebné manuálne vytvoriť štruktúru tabuľky v *Transkribe* a prepísať text, ktorý tabuľka obsahuje. Ak sú rovnaké tabuľky na viacerých stranách, je možné použiť schému pripravenej štruktúry tabuľky na dávkové rozpoznanie ďalších strán s tabuľkami. Ak teda majú viaceré strany rovnakú štruktúru tabuľky alebo šablónu tabuľky, pripraví sa manuálne tabuľka len pri *prvom* výskyte tabuľky a potom sa distribuuje na ďalšie strany pomocou nástrojov *nomacs.* Na transkripciu tabuliek sa najprv vytvoria *textové oblasti* (TR) pre všetky informácie, ktoré *nepatria* do tabuľky. Týka sa to informácií v hornej, spodnej časti alebo po stranách stránky, ktoré zjavne nie sú súčasťou tabuľky, napríklad: čísla strán, čísla riadkov, termíny, akékoľvek iné označenia alebo anotácie. Následne sa vytvoria *textové oblasti* (TR) pre jednotlivé bunky tabuľky, horizontálne a vertikálne čiary a koriguje sa text v bunkách tabuľky na strane. Grafickú schému tabuľky, ohraničenie tabuľky a buniek je možné použiť na ďalšie rovnaké strany s tabuľkami. Bunky sa ohraničujú pomocou volieb v nástroji Cell borders. V projekte SKRIPTOR veľmi uspokojivo zvládol segmentáciu tabuliek a transkripciu textov v bunkách Imrich Nagy.[14]

14 NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 53 – 67.



Obrázok 10 Príklad tabuľky pripravenej na transkripciu (Imrich Nagy, zbierka Koháry_corresp. v *Transkribus* expert client).

Vysvetlivky:

Definované sú textové oblasti (TR), ohraničené sú horizontálne a vertikálne čiara a bunky, segmentované sú riadky textov v bunkách so základnými čiarami a polygónmi riadkov, identifikované sú štrukturálne prvky, záhlavie tabuľky a názvy stĺpcov s názvami prvkov.

MANUÁLNA TRANSKRIPCIA

Na cvičenie vlastného modelu transkripcie *rukopisu* je potrebné pripraviť cvičnú vzorku *manuálne*. Ide o manuálny prepis strán dokumentu, ktorý môže mať tisíce strán, ale na cvičenie prepisujeme manuálne len malú časť. Pred samotným manuálnym prepisom zistíme, či je dokument homogénny, písaný jednou osobou a „jednou rukou" alebo ide o „niekoľko rúk". Podľa toho vyberieme na manuálny prepis dostatočný počet strán. Na cvičenie modelu je potrebné prepísať ručne asi *5 000 až 15 000 slov, teda asi 25 – 75 strán*. Cvičenie modelu transkripcie *tlačených* textov si vyžaduje manuálnu transkripciu podstatne menších cvičných súborov.

V historických textoch, najmä v rukopisoch sa vyskytuje celý rad špeciálnych znakov, ktoré je vhodné vo vedeckej transkripcii zachovať čo najpresnejšie. V nástroji *Transkribus Expert Client* je možné pridávať *špeciálne znaky* a *Unicode (ISO 10646),* ktoré nie sú dostupné na bežnej klávesnici. Virtuálna klávesnica sa nachádza v poli textového editora v spodnej časti okna expertného klienta. Pomocou tlačidla „Upraviť…" je možné pridávať skratky pre často používané znaky a pridávať nové znaky *Unicode*. Ak je potrebné vytvoriť skratku, stačí ju zadať do stĺpca „Skratka" a na pridanie nových znakov *Unicode* použiť zelené tlačidlo plus.

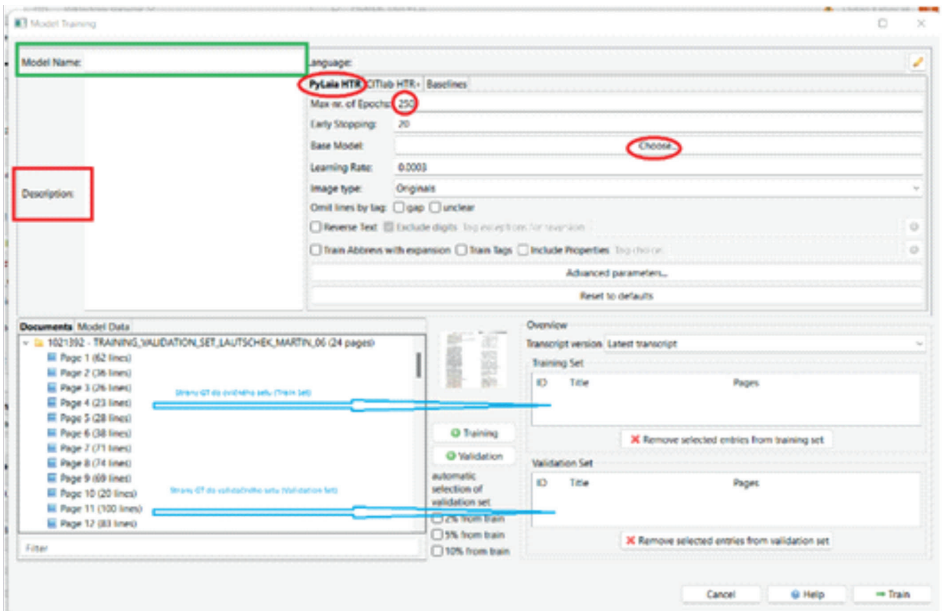
TVORBA MODELU TRANSKRIPCIE

V platforme *Transkribus* je *model* entita, ktorá je výsledkom použitia softvéru strojového učenia a umelej inteligencie a hlbokých neurónových sietí na rozpoznávanie historických rukopisných a tlačených textov. Platforma *Transkribus* umožňuje používateľom cvičiť *model rozpoznávania textu rukou (HTR+, PyLaia)* na automatické spracovanie celej zbierky dokumentov. *Model* je potrebné cvičiť tak, aby rozpoznal určitý štýl písania zobrazovaním obrázkov dokumentov a umožnil ich presný prepis. Podľa typu textu môžu používatelia na transkripciu použiť *verejne dostupný model* alebo vytvoriť vlastný model, prípadne cvičiť vlastný model s použitím základného modelu.

Pred samotným cvičením vlastného modelu zistíme, či na transkripciu nášho dokumentu nebude efektívnejšie použiť nejaký verejný model. Pre každý verejný model je uvedený krátky popis cvičného materiálu, pre ktoré jazyky môže byť model užitočný a kto ho vytvoril a cvičil. Cieľom úsilia komunity *Transkribus* je sprístupniť používateľom *Transkribu* čoraz viac modelov, aby mohli ťažiť z kooperácie a sieťového efektu a šetriť prácu a čas. V novembri 2022 bolo dostupných 97 verejných modelov, napríklad: nemecký kurent, noviny, časopisy, rôzne tlače a rukopisy; viacjazyčný model pre tlače v rôznych jazykoch (holandčina, angličtina, fínčina, francúzština, nemčina, švédčina); všeobecný model pre francúzske rukopisy, nemecká bastarda 15. st.; dánska fraktúra a historické rukopisy a strojopisy; holandské rukopisy a tlače; estónske rukopisy; fínske noviny a rukopisy; francúzske rukopisy a tlače; hlaholika; latinčina; neolatinčina; ruština; španielske rukopisy a tlače a pod.

Pomocou nástroja *Transkribus Expert Client* je môžné cvičiť (trénovať) model rozpoznávania rukopisného textu, aby bolo možné transkribovať automaticky aj rozsiahle zbierky dokumentov. Model je výsledkom cvičenia, preto je pri jeho tvorbe potrebné cvičiť tak, aby stroj rozpoznal určitý štýl písania v zobrazovaných obrázkoch dokumentov a poskytol ich viac-menej presný prepis.

Na cvičenie modelu použijeme voľbu nástroje (Tools). V časti cvičenie modelov (Model training) zvolíme cvičiť nový model (Train a new model). Zobrazí sa okno Model Training. V predvolenom nastavení je vybratý „PyLaia HTR", motor, ktorý nás zaujíma, ako je znázornené na obrázku nižšie.



Obrázok 11 Nastavenie parametrov modelu a výber cvičných a validačných setov.

Vysvetlivky:

Povinne je potrebné prideliť názov modelu (Model name), popis modelu (Description) a jazyk (Language).

Predvolený stroj je PyLaia. Volí sa počet cyklov (Epochs), v ktorých sa stroj učí.

Číslo 20 v riadku Early Stopping 20 znamená, že ak po 20 cykloch (epochách) CER validačnej sady neklesne, tréning sa zastaví.

Riadok Base Model umožňuje vybrať základný model, ktorý už existuje a je dostupný, čo umožňuje zefektívniť učenie.

Číslo 0.0003 v riadku Learning Rate znamená „Rýchlosť učenia“, čo definuje prírastok z jednej epochy do druhej, teda, ako rýchlo bude cvičenie pokračovať. S vyššou hodnotou bude CER klesať rýchlejšie. Čím vyššia je však hodnota, tým vyššie je riziko prehliadnutia detailov.Táto hodnota je adaptívna a upraví sa automaticky. Tréning je však ovplyvnený hodnotou, s ktorou sa začína. Využíva sa predvolené nastavenie.

V okne Documents sú dokumenty a strany, ktoré sú k dispozícii na presun do cvičného setu a validačného setu. Ide o strany, ktoré boli manuálne prepísané a uložené v kvalite Ground Truth (GT). V strednom okne sú voľby Training a Validation, ktorými sa vyberajú strany buď do cvičného setu (Training Set), alebo do validačného súboru (Validation Set). Validačný súbor môže byť napr. 2 %, 5 %, 10 % zo všetkých strán GT v dokumente. Cvičenie sa spustí tlačidlom Train.

HTR+ a PyLaia sú stroje 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 naďalej prehľadávať pomocou plnotextového vyhľadávania v *Transkrib*e.

Stroje HTR+ a PyLaia vychádzajú zo softvéru *TensorFlow* a *PyTorch,* čo je bezplatná softvérová knižnica s otvoreným zdrojovým kódom pre strojové učenie a umelú inteligenciu.[15] TensorFlow slúži na rozpoznávanie obrázkov. Poskytuje štandardný postup, ktorý zahŕňa triedenie pixelov obrázka, získanie vlastností pixelov, trénovanie obrázka, trénovanie modelu a testovanie modelu oproti vstupom. TensorFlow je možné použiť aj na detekciu jazyka, preklad, rozpoznávanie vzorov rukopisu atď. Ich najbežnejšie uplatnenie je v bankách a poisťovniach pri odhaľovaní podvodov. Dá sa použiť v celom rade úloh, ale špeciálne sa zameriava na cvičenie a odvodzovanie hlbokých neurónových sietí. *TensorFlow* bol vyvinutý tímom *Google Brain* na interné použitie Googlom vo výskume a výrobe. Pôvodná verzia bola vydaná pod licenciou *Apache 2.0* v roku 2015. Google vydal aktualizovanú verziu *TensorFlow* s názvom *TensorFlow 2.0* v septembri 2019. *TensorFlow* je možné použiť v širokej škále programovacích jazykov vrátane *Pythonu*, *JavaScriptu, C++* a *Java*. Táto flexibilita sa hodí pre celý rad aplikácií v mnohých rôznych sektoroch. Jednou z aplikácií *TensorFlow* a *PyTorch* je *Transkribus* so strojmi *HTR+ a PyLaia*. PyTorch umožňuje vyspelejšie rozpoznávanie textu. Na trénovanie modelu rozpoznávania textu založeného na umelej inteligencii sa používa rekurentná neurónová sieť (RNN) a PyTorch. Ďalšie podobné aplikácie založené zahŕňajú detekciu rukopisu, rozpoznávanie vzorov atď.

Ak tvoríme vlastné, generické modely HTR, tak nepracujeme so základnými modelmi. Avšak aj pri cvičení vlastného modelu môže byť cvičenie založené na existujúcom *základnom modeli*, ktorým býva spravidla posledný najlepší model HTR, ktorý bol vyškolený v nejakom projekte. Základné modely si „pamätajú“ to, čo sa už „naučili“. Preto každé nové školenie sa „teoreticky“ zlepšuje kvalitu novovytvoreného modelu. Nový model sa učí od svojho predchodcu a stáva sa tak lepším a lepším. Preto je školenie so základnými modelmi obzvlášť vhodné aj pre veľké generické modely, ktoré sa neustále vyvíjajú počas dlhého časového obdobia. Ak chceme vykonať školenie so základným modelom, jednoducho si v cvičnom nástroji vyberieme konkrétny základný model – okrem obvyklých nastavení. Potom na karte *Údaje modelu HTR* vložíme cvičný súbor a overovací súbor základného modelu, ako aj nový cvičný a overovací súbor. Okrem toho môžeme pridať ďalšie nové strany *Ground Truth* a začať s cvičením.

Pri práci so systémom *Transkribus Expert Client* sa pri každom spustení úlohy alebo uložení dokumentu vytvorí nová verzia dokumentu. Výhodou je, že sa vždy môžete vrátiť k starším verziám a pokračovať v práci na nich, čo sťažuje stratu údajov v *Transkribe*. Navyše je možné porovnávať verzie s nástrojom *Compute Accuracy v Transkribe*. Pri verziách jednotlivých stránok je vždy informácia o stave (statuse) strany, používateľa, dátum zmeny, nástroj zmeny a identifikátory.

Presnosť modelu je možné merať na konkrétnych stránkach z cvičných a overovacích

15 JAISWAL, Abhishek: PyTorch vs TensorFlow: What is Best for Deep Learning in 2023? In: *Turing* [online]. [cit. 2022-11-20]. Dostupné na: https://www.turing.com/kb/pytorch-vs-tensorflow

súborov pomocou funkcie *„Presnosť výpočtu“* (Compute accuracy...) na karte „Nástroje“ (Tools). Na tento účel je najprv potrebné vygenerovať transkripciu HTR. Na porovnanie textových verzií sú potrebné dva transkribované súbory: *„Referencia“* (Reference) – správny text a *„Hypotéza“* (HTR, transkribovaný text). Ako **„Referencia“** sa vyberie verzia stránky, ktorá bola správne prepísaná, teda „Základná pravda“ (Ground Truth), čo je manuálny prepis čo najbližšie k pôvodnému textu. Na získanie najvýznamnejšej hodnoty by bolo najlepšie použiť stránky zo vzorového súboru, ktoré neboli použité v tréningu, a preto sú pre model nové. Použitie stránok z overovacieho súboru po úpravách je tiež možnosťou, aj keď nie je taká ideálna. Použitie stránok z cvičného súboru nie je vhodné, pretože to prinesie nižšie hodnoty CER než v skutočnosti sú. Ako ***„Hypotézu“*** vyberieme verziu, ktorá bola automaticky vygenerovaná pomocou modelu HTR, na ktorej chceme vidieť, aký dobrý je výsledok.

CER (Character Error Rate) je miera chybovosti znakov. Porovnáva sa pre danú stranu celkový počet znakov (n) vrátane medzier s minimálnym počtom vložení (i), nahradenia (s) a vymazania (d) znakov, ktoré sú potrebné na získanie výsledku *Ground Truth*. Ide teda o chyby v porovnaní s presným, referenčným textom. Vzorec na výpočet CER je takýto: 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 chyby. Považuje za potvrdené a overené konštatovanie, že: a) Ak je hodnota chybovosti *znakov* CER 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*.

Tabuľka 1 Empirické poznatky platformy *Transkribus* o korelácii chybovosti znakov a cvičných dát[16].

	CER (Character Error Rate)	**Cvičné dáta (Train)**
Tlačený text	0,5-2%	~ 5.000 slov
Jednoduchý rukopis – jedna ruka	2-4%	+10.000 slov
Niekoľko rúk – všetky vidieť počas tréningu	4-6%	10 000 slov (špecificky podľa rúk)
Veľa rúk z rovnakého obdobia a regiónu – nie všetky vidieť počas tréningu	6-8%	+100.000 slov (nešpecifikované ruky)

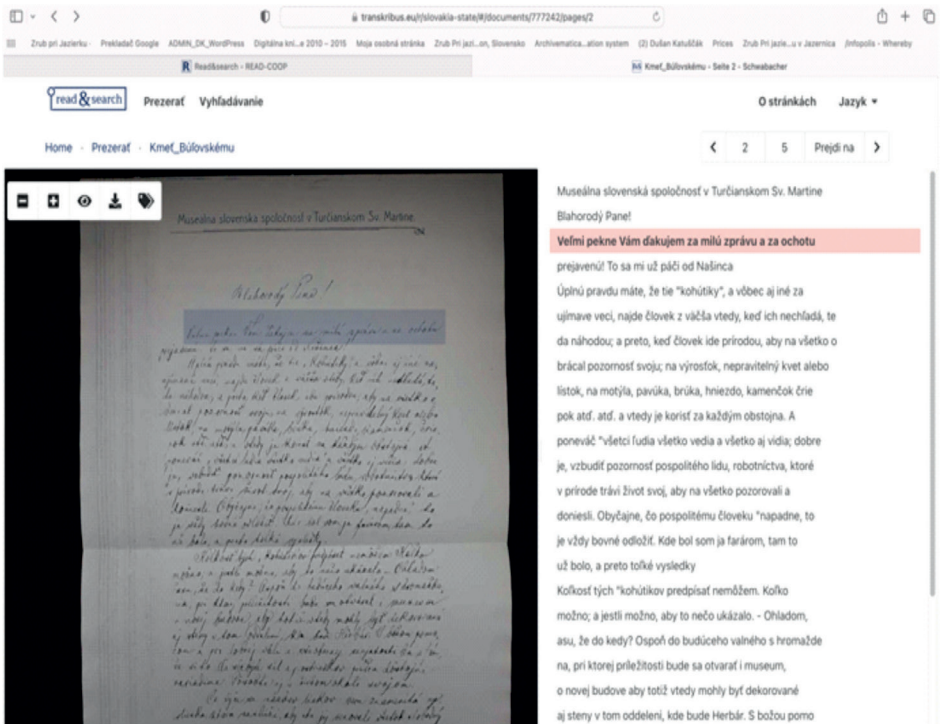
SPRÍSTUPNENIE A POUŽITIE VÝSLEDKOV TRANSKRIPCIE

Sprístupnenie dokumentov a výsledkov transkripcie umožňuje nástroj *read&search,* ďalej digitálny repozitár *DSpace Univerzity Mateja Bela* a digitálny repozitár *Tainacan* v Štátnej vedeckej knižnici v Banskej Bystrici.

16 How To Train and Apply Handwritten Text Recognition Models in Transkribus eXpert. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-11-20]. Dostupné na: https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/

Ak chceme pracovať so svojimi obrázkami a prepismi mimo *Transkribu*, môžeme svoje dokumenty exportovať do bežnejších formátov, ako sú DOCX, PDF, EXCEL, XML, TEI-XML alebo TXT. Možnosti zahŕňajú *export* celých strán, obrázkov, textu alebo štrukturálnych prvkov. Exportovať je možné do nejakého adresára na lokálnom počítači alebo exportovať na server *Transkribus*, z ktorého nám príde oznámenie po skončení exportu. *Editovať a korigovať* výsledky transkripcie je možné aj so zapojením dobrovoľníkov a odborníkov, napríklad aj v *Transkribus Lite*, čo je verzia prehliadača *Transkribus*. Automaticky transkribuje a umožňuje pohodlnú úpravu historických dokumentov. V *Transkribus Lite* je tiež možné cvičiť vlastné modely AI v nejakom prehliadači. Tu, v prehliadačoch osobných počítačov a smartfónov je možné prezerať a upravovať zbierky z *Transkribus Expert Client*. Mnohé z funkcií klienta *Transkribus Expert Client* môžu byť použité aj v *Transkribus Lite*. V *Transkribus Lite* sa analýza rozloženia (segmentácia) spustí automaticky, keď sa spustí úloha rozpoznávania textu, a nie je možné ju upravovať. Limitovaný tu môže byť aj menší počet strán na transkripciu; v súčasnosti je limit 10 MB.

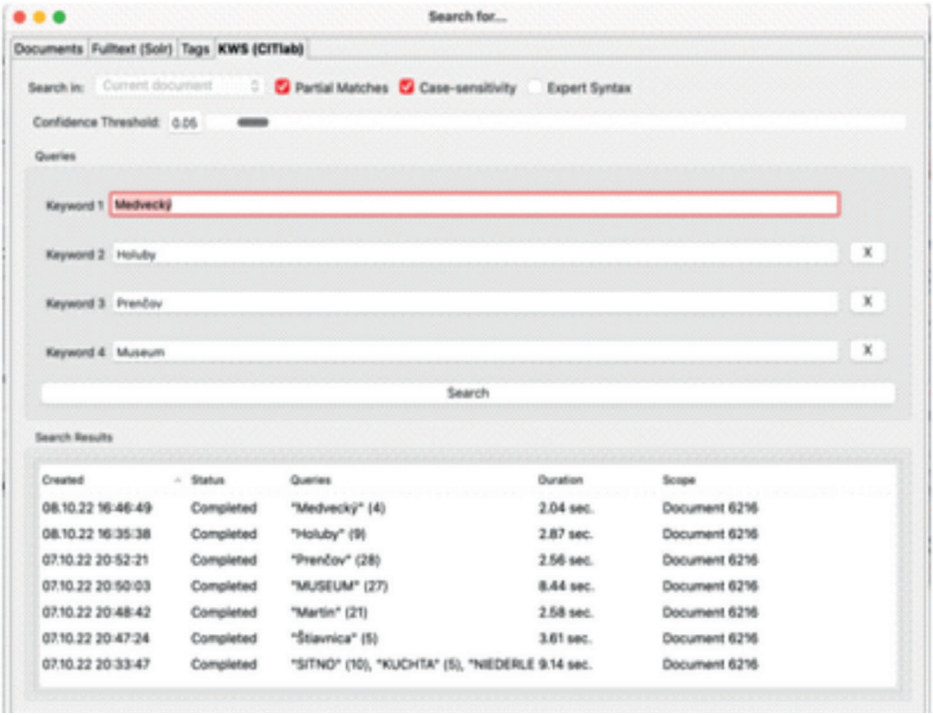
Read&search je platená platforma *Transkribu,* ktorý sprístupňuje online dokumenty zo zbierky vytvorenej v platforme *Transkribus Expert Client*. Toto rozhranie bohaté na funkcie je ideálne na sprístupnenie historických dokumentov a vyhľadávanie na webe. Alternatívne je možné po exporte výstupy transkripcie a zdrojové obrázky sprístupňovať aj na iných webových sídlach alebo z digitálnych repozitárov (DSpace, Greenstone, Tainacan, MediaInfo a i.).



Obrázok 12 Sprístupnenie výsledku automatickej transkripcie v read&search.

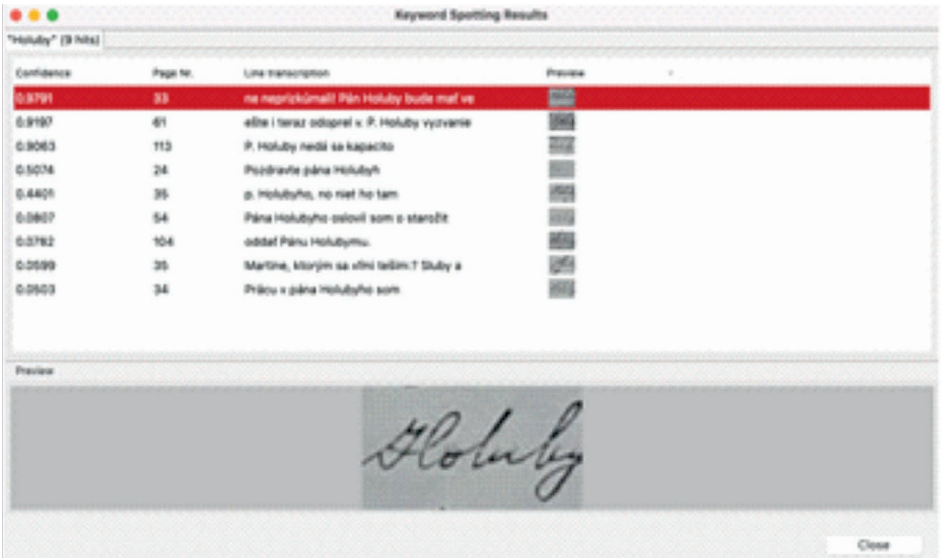
VYHĽADÁVANIE

V dokumentoch, ktoré boli v *Transkribe* transkribované pomocou HTR-modelu, je možné *vyhľadávať* pomocou kľúčových slov použitím fulltextového vyhľadávania (Solr). Systém umožňuje (pravdepodobnostné) „fuzzy vyhľadávanie“ (Fuzzy Search), čo je vyhľadávacia technika, ktorá umožňuje nájsť podobné slová nielen podľa presnej zhody s hľadaným výrazom. Pomocou expertného klienta je možné zadať „fuzzy vyhľadávanie“, čo znamená, že nie všetky znaky hľadaného reťazca znakov sa musia zhodovať. Text by mal byť okamžite po hľadaní k dispozícii. Indexuje sa vždy iba posledná verzia každého prepisu. Vyhľadávanie je možné po zadaní jedného slova, viacerých slov alebo presnej vety. Vyhľadávanie je možné aj podľa tagov (značiek) v dokumentoch v zbierke, v dokumente, na strane, riadku, s voľbou veľkých a malých písem. Originálne vyhľadávanie umožňuje *metóda KWS* (The Keyword Spotting). V zbierke je možné hľadať podľa ID zbierky, ID dokumentu, názvu zbierky, popisu a autora zbierky. Zaujímavý nástroj na vyhľadávanie je KWS, 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.



Obrázok 13 Vyhľadávanie pomocou nástroja KWS (hľadané slová a ich výskyty v rukopise).

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 0 a 1 (0 = najnižšia a 1 = najvyššia), aby sa zhodnotila kvalita výsledkov hľadania.



Obrázok 14 Použitie KWS, vyhľadané obrazy slov a ich výskyty.

PREKLAD

Výsledný transkribovaný a exportovaný text môže byť preložený z jazyka transkripcie do iného jazyka pomocou strojového prekladu. Možno, azda oprávnene, predpokladať, že následný preklad transkribovaných textov nebude možný okamžite, ale bude si vyžadovať rovnaké úsilie ako transkripcia. Automatické prekladače sa budú musieť učiť prekladať historické texty a rovnako budú potrebovať na učenie veľké súbory cvičných dát.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

BEŽO, Ján – GORÍŠEK, Karol: *Šlabikár a Prvá čítanka pre školy evanjelické a. v*. Senica : Nákladom Jána Bežo, učiteľa v Senici, 1879, s. [25]. Dostupné na: https://onk.snk.sk/view/986df890-9133-48bf-8c25-cda6cee3c724?page=9476cad8-544e-44ad-a6b4-450a0f6397db

DIETRICH, Felix: OCR vs. HTR or "What is AI, actually?" In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-11-20]. Dostupné na: https://readcoop.eu/insights/ocr-vs-htr/

DRAŠKABA, Peter – HANUS, Jozef: Všeobecná medzinárodná norma pre opis archívnej jednotky. In: *Slovenská archivistika,* roč. 35, č. 1, 2000, s. 197 – 215.

Glossary. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-12-09]. Dostupné: https://readcoop.eu/glossary/

How To Train and Apply Handwritten Text Recognition Models in Transkribus eXpert. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-11-20]. Dostupné na: https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/

International Council on Archives: *Archival Arrangement & Description : Global Practices. Report on the survey undertaken by the ICA Training Programme, with a foreword by the ICA President* [online]. July 2020 [cit. 2023-01-20]. Dostupné na: https://www.ica.org/sites/default/files/aad_survey_report_final_202108_eng.pdf

ISAD(G) : general International Standard Archival Description : adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 September 1999 [online]. Ottawa, 2000 [cit. 2022-11-20]. Dostupné na: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

ISAD(G) : všeobecný mezinárodní standard pro archivní popis. Přijato Komisí pro popisné standardy, Stockholm, Švédsko, 19. – 22. září 1999 [online]. Praha, 2009, 57 s. [cit. 2022-11-20]. Dostupné na: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

JAISWAL, Abhishek: PyTorch vs TensorFlow: What is Best for Deep Learning in 2023? In: *Turing* [online]. [cit. 2022-11-20]. Dostupné na: https://www.turing.com/kb/pytorch-vs-tensorflow

LIŠKA, Matej: *Zoner Photo Studio X. Praktická príručka*. Brno : Zoner software, 2021. 178 s.

MARETTA, Gregor Robert: Digitalizácia stredovekých listín v Slovenskom národnom archíve. In: *Slovenská archivistika*, roč. 34, č. 1, 2009, s. 16 – 40.

MUEHLBERGER, Guenter et al.: Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: *Journal of Documentation* [online], vol. 75, no. 5, 2019, pp. 954 – 976 [cit. 2021-10-06]. Dostupné na: https://doi.org/10.1108/JD-07-2018-0114

MÜHLBERGER, Günter: *READ (Recognition and Enrichment of Archival Documents) – 2016 – 2019* [online]. [cit 2021-10-06]. Dostupné na: https://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019

NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 53 – 67.

ORMIS, Samuel: *Ewanjelický Šlabikář wypracowaný z naľožení wel. Seniorátu gemerského*. V Rožnavě : Tiskem Michala Kovács, 1872, s. [1]. Dostupné na: https://onk.snk.sk/view/c9e7ce09-2c3b-4c72-85e9-304947366228?page=f7db845e-e475-460b-9c69-cf03a3f9806e

Oznámenie generálneho riaditeľa sekcie verejnej správy o uverejnení opatrenia ministra vnútra Slovenskej republiky č. SVS-204-2008/00111 zo 17. januára 2009 o správnych informáciách a službách štátnych archívov zriadených Ministerstvom vnútra Slovenskej republiky, čl. 10. In: *Vestník Ministerstva vnútra Slovenskej republiky*, roč. 2009, čiastka 6, 29. január 2009.

POPOVIČ, Anton: *Originál – preklad. Interpretačná terminológia*. Bratislava : Tatran, 1983. 362 s.

Pravidlá slovenského pravopisu. red. M. Považaj. 4. nezm. vydanie, Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 2013. 592 s.

READ Recognition and Enrichment of Archival Documents. In: *CORDIS: EU research results* [online]. Last update 17 August 2022 [cit. 2022-11-20]. Dostupné na: https://cordis.europa.eu/project/id/674943

Resource center. In: *READ-COOP* [online]. Innsbruck: READ-COOP SCE, 2021 [cit. 2022-11-20]. Dostupné na: https://readcoop.eu/transkribus/resources/

KAPITOLA 2

POSTILA IZÁKA ABRAHAMIDESA HROCHOTSKÉHO A AUTOMATICKÉ ROZPOZNÁVANIE JEHO RUKOPISU

Pavol Maliniak

Univerzita Mateja Bela v Banskej Bystrici; Filozofická fakulta; Katedra histórie
E-mail: pavol.maliniak@umb.sk

ABSTRAKT

Izák Abrahamides Hrochotský, evanjelický kazateľ vo Zvolene, je autorom zbierky kázní z rokov 1600 – 1601. V originálnom rukopise písanom v slovakizovanej češtine sa strieda novogotické a humanistické písmo. Pri trénovaní modelov automatickej transkripcie boli zvolené vzorky manuálneho prepisu v ortograficky vernej podobe bez rozpisovania skratiek. Doposiaľ boli trénované tri modely strojového učenia s hodnotou Ground Truth. Tretí model s rozsahom 11 434 slov je najúspešnejší. Chybovosť znakov CER v cvičnom súbore predstavovala 1,55 %. V overovacom súbore bola hodnota CER 7,99 %, čo je použiteľný výsledok. Otázny je vzťah medzi chybovosťou a rôznymi štýlmi písma. Pre automatický prepis bola preto vybraná strana rukopisu, kde sa vyskytuje iba text v latinskom jazyku. Chybovosť prepisu bola až 12,83 %. Vysoký podiel chýb ukázala najmä humanistická majuskula.

Kľúčové slová: rukopisné kázne; viacjazyčný text; Izák Abrahamides Hrochotský; cvičenie modelov; miera chybovosti znakov; HTR+; *Transkribus*; transliterácia

ABSTRACT

Isaac Abrahamides Hrochotius' postil and automatic recognition of his handwriting

Isaac Abrahamides Hrochotius, an evangelical preacher in Zvolen, is the author of a collection of sermons from the years 1600-1601. The original manuscript, written in Slovakized Czech, alternates between Neo-Gothic and humanistic script. When training automatic transcription models, manual transcription samples were selected in an orthographically faithful form without breaking down abbreviations. So far, three machine learning models have been trained with the Ground Truth value. The third model with the range of 11 434 words is the most successful one. The error rate of CER characters in the training set was 1.55%. In the validation file, the CER value was 7.99%, which is a usable result. The relationship between error rate and different font styles is questionable. Therefore, the manuscript's page selected for the automatic transcription was one with text solely in Latin. The transcription error rate was up to 12.83%. A high share was shown especially by the humanistic capital lettering.

Key words: manuscript sermons; multilingual text; Isaac Abrahamides Hrochotius; model training; character error rate; HTR+; *Transkribus*; transliteration

ÚVOD

S obdobím reformácie a renesancie v Uhorsku je spojená osobnosť Izáka Abrahamidesa (s prímenom Hrochotský), ktorý je autorom pozoruhodného, avšak doposiaľ nedostatočne reflektovaného diela. Počas účinkovania vo funkcii evanjelického farára vo Zvolene na rozhraní 16. a 17. storočia napísal Abrahamides zbierku kázní, ku ktorej je podľa súčasných poznatkov známych iba málo analógií. Vo vzťahu k digitálnym humanitným vedám je prednosťou prameňa jeho vznik v krátkom časovom úseku a relatívne veľký rozsah. Abrahamidesova postila je v širšom kontexte dôležitým zdrojom poznania náboženských a cirkevných dejín, rovnako aj dejín mentalít a každodenného života alebo dejín literatúry a jazykových pomerov. Poskytuje preto rôznorodé námety pre medziodborové výskumy. Prameň môže okrem historiografie využiť najmä bádanie z oblasti filológie, teológie, religionistiky, etnológie, ale aj kodikológie.

AUTOR POSTILY

Základom životopisných údajov o Izákovi Abrahamidesovi bola dlhodobo zbierka historických dokumentov k dejinám reformácie *Collectanea* (XXII. zväzok) zostavená v závere 18. storočia Martinom Laučekom. Dôležité pasáže z nej uverejnil Jozef Ľudovít Holuby.[1] Laučekove údaje čiastočne vychádzali zo životopisných dát, ktoré uviedol samotný Abrahamides, keď bol vo Wittenbergu v roku 1595 ordinovaný za kňaza.[2] Doposiaľ zrejme najkomplexnejšie spracovanú biografiu predstavil dobronivský evanjelický farár Ján Slávik vo svojich dejinách Zvolenského seniorátu. Z hľadiska novších výskumov si zasluhujú pozornosť zistenia Denisa Pongrácza. Na základe genealogického aj heraldického výskumu spresnil Abrahamidesov pôvod – spojil ho presvedčivo s rodinou nobilitovaného zvolenského mešťana Abraháma (s prímenom Bartolen) a jeho manželky Anny Churhayovej (Čurhaj). Z postupne precizovaných údajov možno zhrnúť, že Izák Abrahamides Hrochotský sa narodil v roku 1557. Základy vzdelania nadobudol v mestskej škole vo Zvolene. Aby si osvojil jazyky, študoval štyri roky v Banskej Štiavnici, dva roky v Banskej Bystrici a dva roky v Bardejove. Tri roky sa učil gréčtinu a hebrejčinu v Mošovciach. Hodnosť magistra filozofie dosiahol po dvoch rokoch štúdia na univerzite v Prahe. Pokračoval ešte rok v štúdiu na univerzite v Lipsku. Potom štyri roky účinkoval ako vychovávateľ mladých uhorských barónov vo Viedni. Tri roky strávil na cisárskom dvore (*in aula imperatoria*) Rudolfa II. Asi v roku 1586 sa stal rektorom školy vo Zvolene. Od roku 1590 zastával funkciu mestského notára v Kremnici, pričom zastupoval aj okolité

1 HOLUBY, Jozef Ľudovít: Životopis superintendenta Mr. Izáka Abrahamidesa. Z pôvodného rukopisu Laučekovho, r. 1795 písaného, a v Zay-Uhrovskom archíve opatrovaného. In: *Cirkevné listy*, roč. XXIV, č. 4, 1910, s. 114 – 115; č. 6, 1910, s. 171 – 175.

2 BUCHWALD, Georg: Beiträge zur Kenntniss der evangelischen Geistlichen und Lehrer Oesterreichs aus den Wittenberger Ordiniertenbüchern seit dem Jahre 1573. In: *Jahrbuch der Gesellschaft für die Geschichte des Protestantismus in Oesterreich* [online]. 21. Jg., 1900, s. 123 – 124 [cit. 2022-11-21]. Dostupné na: https://anno.onb.ac.at/cgi-content/anno-plus?aid=jgp&datum=1900&page=117&size=45

banské mestá ako vyslanec na panovníckom dvore. O výraznejšej zmene v jeho živote možno hovoriť od roku 1595, keď nastúpil dráhu evanjelického duchovného vo Zvolene. V meste pracoval do roku 1607, keď odišiel do Bojníc, kde pôsobil v úrade prepošta. Na Žilinskej synode v roku 1610 bol Abrahamides v neprítomnosti zvolený za superintendenta pre Nitriansku, Tekovskú a Bratislavskú stolicu. Na pohrebe palatína Juraja Thurza v Bytči v roku 1617 predniesol v latinčine pohrebnú kázeň *Oratio Exequialis*, vydanú následne tlačou.[3] Izák Abrahamides zomrel v Bojniciach 30. augusta 1621. Vdove Apolónii Reichardtovej zanechal rozsiahlu knižnicu, o ktorú vzápätí prejavoval záujem Stanislav Thurzo.[4] Podľa ďalších zistení pôsobili v kňazskej službe prinajmenšom dve generácie jeho potomkov. Izákov syn Šimon Abrahamides bol evanjelickým duchovným v Jelšave, Novom Meste nad Váhom a v Skalici. Rovnako dosiahol funkciu prepošta. Šimonov syn Daniel Abrahamides bol dvorným kazateľom rodu Sibrikovcov v Bozsoku (v Zadunajsku) a maďarsko-slovenským kazateľom v Bratislave.[5] Nemožno vylúčiť, že skúmané rukopisné kázne spočiatku vlastnili a používali obaja potomkovia Izáka Abrahamidesa.

OSUDY RUKOPISU

Existenciu rukopisnej postily bádanie dokumentuje až od 19. storočia, keď bola vo vlastníctve peštianskeho evanjelického farára a literáta Jána Kollára. V prvom zväzku svojho diela *Národnié Zpiewanky* v roku 1834 uviedol, že sa u neho nachádza originál kázní Izáka Abrahamidesa a mieni ich vydať tlačou.[6] Plán vydať kázne Kollár nerealizoval. Rukopis zostal v knižnici slovenskej evanjelickej cirkvi v Budapešti, kde sa nachádzal ešte v roku 1880. Podľa dedukcií Fedora Ruppeldta, ktorý štúdiu postily doposiaľ venoval najväčšiu pozornosť, sa rukopis mal údajne dostať do vlastníctva Sama Chalupku. Po jeho smrti v roku 1883 mal dokument získať vydavateľ jeho diela, banskobystrický advokát Ľudovít Turzo Nosický. Doposiaľ publikovaná korešpondencia Sama Chalupku však neobsahuje

3 *Oratio Exequialis, Illustrissimo p. m. Comiti ac Domino, D. Georgio Thvrzoni de Bethlemfalwa* [online]. Dicta et Recitata ab ISAACO Abrahamide HROCHOTIO, Levtschoviae : Daniel Schultz, 1617 [cit. 2022-11-18]. Dostupné na: http://147.213.131.4:85/digi/mf/Mf_084/start.htm

4 BUCHWALD, G.: Beiträge zur Kenntniss, s. 123 – 124; SLÁVIK, Ján: Vysvätenie Izáka Abrahamidesa Hrochoťského na biskupstvo. In: *Cirkevné listy*, roč. XXXIII, č. 4 – 5, 1919, s. 104 – 108; SLÁVIK, Ján: *Dejiny zvolenského evanjelického a. v. bratstva a seniorátu.* Banská Štiavnica : Tlačou a nákladom vdovy a syna Augusta Joergesa, 1921, s. 799 – 800; ČAPLOVIČ, Ján: Z osudov starších slovenských súkromných knižníc. In: *Knižnica : časopis pre knižnú kultúru*, roč. III – IV, 1951 – 2, s. 41 – 42; RUPPELDT, Fedor: Izák Abrahamides – kazateľ. In: *Cirkevné listy*, roč. 76, č. 7 – 8, 1963, s. 124; ŽILÁK, Ondrej: Izák Abrahamides. In: *Cirkevné listy*, roč. 84, č. 9, 1971, s. 136 – 138; KOVAČKA, Miloš et al.: Osobnosti Žilinskej synody. Malý biobibliografický slovník. In: *Akty a závery – zákony a ustanovenia Žilinskej synody.* ed. M. Kovačka, Martin : Slovenská národná knižnica, 2010, s. 66 – 67; PONGRÁCZ, Denis: *Atlas osobných pečatí I.* Bratislava : Vydal JUDr. Mikuláš Trstenský vlastným nákladom, 2019, s. 10.

5 CSEPREGI, Zoltán: *Evangélikus lelkészek Magyarországon (ELEM) II/1. A zsolnai zsinattól (1610) a soproni országgyűlésig (1681). II/1: Nyugat-Magyarország (a dunántúli, a bajmóci és a felső-dunamelléki egyházkerület)* [online]. Budapest : RECITI, 2020, s. 43 [cit. 2022-11-21]. Dostupné na: https://mek.oszk.hu/23000/23048/23048.pdf

6 RUPPELDT, F. Izák Abrahamides – kazateľ, s. 124 – 125; SLÁVIK, J. *Dejiny zvolenského evanjelického...*, s. 800.

údaj o tom, že prejavoval záujem o zbierku Abrahamidesových kázní, alebo že sa nachádzala v jeho vlastníctve. Argumentácia F. Ruppeldta vychádzala z neskoršej situácie, pretože spolu s postilou sa v súbore kníh a písomností zachovali aj listy Boženy Němcovej adresované S. Chalupkovi. Z Turzovej pozostalosti sa postila mala dostať do rúk banskobystrickej rodiny Bothárovcov. Za kuriózny, hoci v slovenských pomeroch nie ojedinelý, možno označiť údaj, podľa ktorého rukopis kázní spolu s inými dokumentmi bol určený na spálenie. Písomnosti mal objaviť Daniel Bothár v drevárni u svojho brata počas návštevy Banskej Bystrice v roku 1924. D. Bothár rukopisy previezol do Šopronu, kde pôsobil ako gymnaziálny profesor. Po jeho smrti v roku 1929 zdedil postilu spolu s ďalšími slovacikami jeho syn Michal Bothár, farár v rakúskom Schlainingu. Ešte v tom istom roku celý súbor slovacík vo Viedni od neho odkúpil za 10 000 korún Fedor Ruppeldt. Dokumenty previezol do Žiliny, kde pôsobil ako evanjelický farár a neskôr biskup.[7] Samotný rukopis postily začal Ruppeldt študovať až po odchode do penzie od konca 50. rokov. V roku 1963 publikoval v *Cirkevných listoch* prvú podrobnejšiu charakteristiku prameňa, ktorý už bádanie považovalo za stratený. Okrem toho zanechal niekoľko desiatok strán poznámok, komentárov a postrehov. Postilu najprv stručne opísal po formálnej stránke. Zvýšenú pozornosť venoval jazyku a pravopisu, s dôrazom na výskyt slovakizmov, vplyv iných jazykov, ľudové výrazy a príslovia. Zameral sa i na zhodnotenie obsahu kázní z homiletického, teologického a čiastočne aj historického hľadiska. Napokon sa sústredil na osobnosť Izáka Abrahamidesa a dobu, v ktorej kázne vznikli.[8] Úvahy sa spolu s originálom postily dostali v osobnom fonde F. Ruppeldta do Literárneho archívu Slovenskej národnej knižnice v Martine. Na prítomnosť rukopisu a bohaté možnosti jeho štúdia v nedávnom období opäť upozornil teológ Igor Kišš (pre prameň navrhol pomenovanie Zvolenská postila) v niekoľkých článkoch pri príležitosti jubilea Žilinskej synody. Postilu zároveň pracovníci knižnice, resp. archívu zdigitalizovali a poskytujú ju pre bádanie, zatiaľ bez online prístupu.[9]

PÍSMO A JAZYK

Postila Izáka Abrahamidesa zachovaná v origináli má rozsah 722 strán, resp. pagín. Tvorí ju husto popísaný text malého formátu (približne 15,6 x 10,7 cm), pričom niekoľko desiatok strán je prázdnych alebo sčasti zaplnených. Stav rukopisu je navzdory jeho osudom vyhovujúci, pretože porušené s čiastočne chýbajúcim textom sú iba dva listy (štyri strany). Papierové listy sú zviazané do zväzku bez samostatného obalu s hrúbkou približne 6 cm.[10] Rukopis je značený staršou

7 Slovenská národná knižnica – Literárny archív (ďalej LA SNK), Zbierky a fondy, Literárne rukopisy, XXXI, 240 Fedor Ruppeldt, sign. 240 AN 3, *Úvahy o kázňovke – rukopise I. Abrahamidesa z r. 1600 – 1601*, kapitola 4, s. 8 – 10, 13 – 15; RUPPELDT, F.: Izák Abrahamides – kazateľ, s. 124 – 125; KIŠŠ, Igor: Výskumy Fedora Ruppeldta na Abrahamidesovej postile z r. 1600 – 1601. In: *Cirkevné listy*, roč. 134, č. 5, 2010, s. 10 – 11; ORMIS, Ján V. (ed.): Listy Sama Chalupku. In: *Litteraria. Štúdie a dokumenty II.* red. O. Čepan, Bratislava : Vydavateľstvo Slovenskej akadémie vied, 1959, s. 93 – 149.

8 RUPPELDT, F.: Izák Abrahamides – kazateľ, s. 125 – 126.

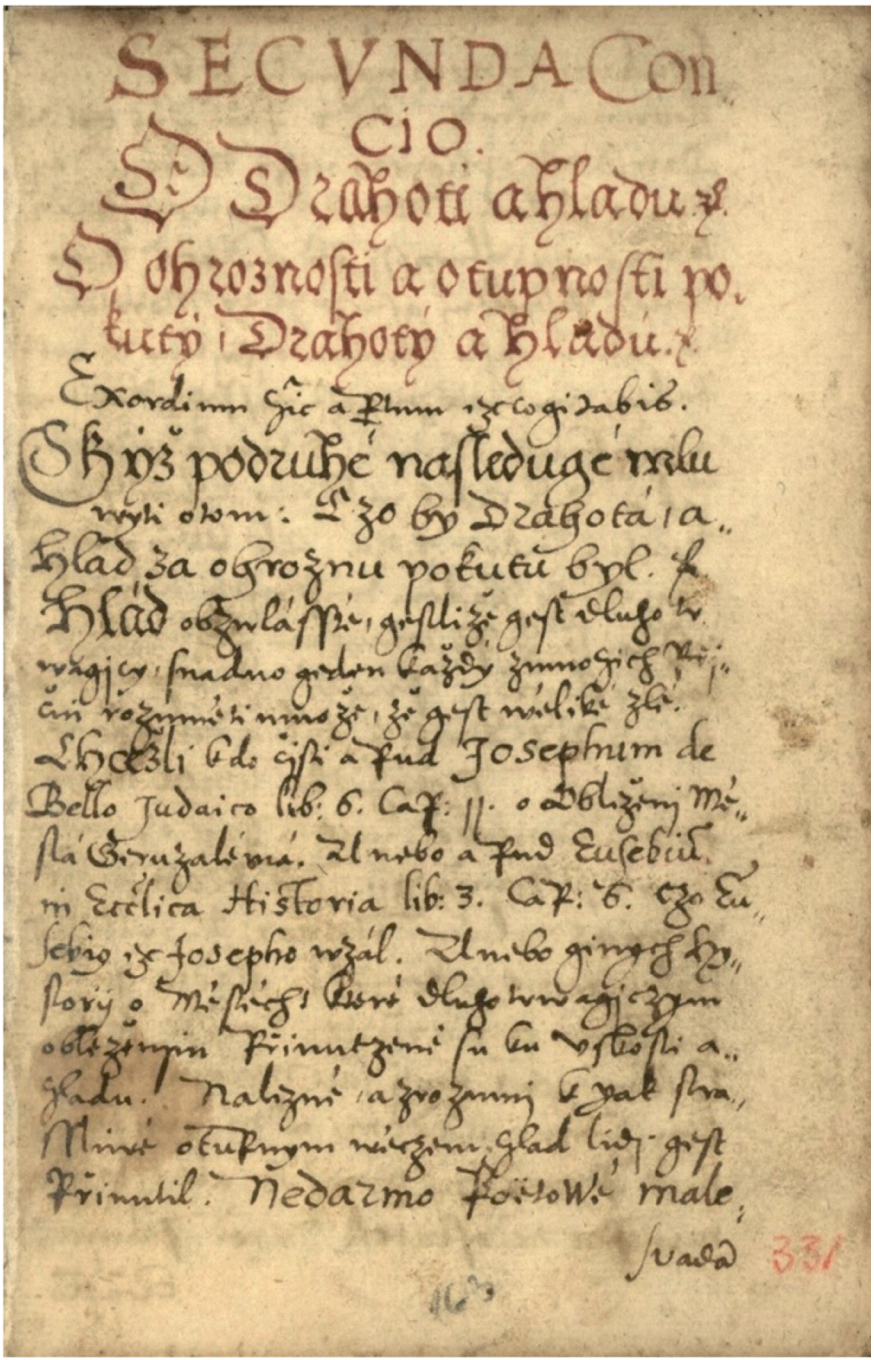
9 LA SNK, sign. 240 AN 3, *Úvahy o kázňovke – rukopise I. Abrahamidesa*, 11 kapitol a nečíslované strany; KIŠŠ, Igor: Výskumy Fedora Ruppeldta, s. 9; KIŠŠ, Igor: Úsilie Izáka Abrahamidesa o výchovu k humánnej spoločnosti. In: *Slovenské pohľady*, roč. IV. + 129, č. 5, 2013, s. 63 – 64.

10 LA SNK, sign. 240 BN 1, pag. 273 – 274, 721 – 722; KALAJTZIDIS, Ján: Superintendent Izák Abrahamides Hrochotský a jeho Zvolenská postila. In: *Historia Ecclesiastica*, roč. XI, č. 2, 2020, s. 71.

paginációu a mladšou foliáciou, v obidvoch prípadoch nepresne. V rámci paginácie bolo vynechané číslovanie pag. 324 – 325 a namiesto neho bolo uvedené číslovanie 326 – 327 s nadväzujúcim posunutým značením. V prípade foliácie sú duplicitne číslované listy fol. 28r-v, pričom druhýkrát malo ísť už o 29r-v. Na porušenom liste fol. 136r-v číslovanie chýba, pričom ako 136r-v je označený nasledujúci list, opäť s nadväzujúcim posunutým číslovaním. Doterajšie bádanie uprednostňovalo pri citovaní prameňa paginációu značenú kontrastnejšími červenými číslicami (na rozdiel od foliácie). Tento postup možno s vedomím určitej nepresnosti využívať aj pri súčasnom výskume. Digitalizácia a s ňou spojená interpretácia rukopisu zároveň vytvárajú možnosti na presnejšie označovanie a citovanie jednotlivých pagín alebo folií.

Rukopis možno považovať za čistopis s minimálnymi neskoršími zásahmi, najmä podčiarkovanými slovami a marginálnymi poznámkami (skratky biblických kníh) červenou a čiernou ceruzou. Súvislý text napísaný mladšou rukou, azda z 19. storočia, s iniciálkami J. M., sa nachádza iba na titulnej strane zväzku. Ide o memoriálny zápis prisudzujúci autorstvo postily Izákovi Abrahamidesovi.[11] Nasledujúci text kázní je písaný úhľadne jednou písarskou rukou, zahŕňa však rôzne štýly a druhy písma pokročilého 16. storočia. S výnimkou niekoľkých slov v gréckom a hebrejskom písme rukopis pozostáva prevažne z novogotického, v menšom rozsahu aj z humanistického písma. Na použitý štýl, ale aj na hustotu, rozloženie a riadkovanie vplýva použitý jazyk. V češtine s rôznym zastúpením slovakizmov je písaný text v novogotickej kurzíve s charakteristickým sklonom, dlhými driekmi v literách *b*, *d*, *f*, *j*, *k*, *l*, *s*, *t* a ligatúrami najmä medzi spoluhláskami. Horné a dolné dĺžky litier často presahujú do susedných riadkov. Predovšetkým majuskulné litery nadobúdajú zložité tvary. Uplatnenie majuskuly v slovách a takisto aj interpunkcie a diakritiky je značne nedôsledné, ako je to obvyklé pre ranonovoveký pisársky úzus. Na pasáže v latinčine je viazané jednoduchšie humanistické písmo, ktoré je síce čitateľnejšie, ale v celom rukopise menej zastúpené. V prípade kurzívy nie je ortografia v latinskom jazyku ustálená, ale možno hovoriť aj o humanistickej kurzíve s vplyvmi kurentu. Vzhľadom na liturgické zameranie rukopisu a vzdelanie pisára je popri novogotickej kurzíve bežne používaná i fraktúra, ktorou sú ozdobované zvlášť nadpisy, úvody a slová zdôraznené v ostatnom texte. Podobne v latinských pasážach sa popri kurzíve možno stretnúť v nadpisoch a zvýraznených slovách s rovnomerne rozostavanou humanistickou majuskulou. V postile sa opakovane vyskytujú strany, kde sa striedajú štyri rukopisné prejavy: humanistická majuskula, novogotické kreslené písmo, novogotická a humanistická kurzíva (obrázok 15).

11 LA SNK, sign. 240 BN 1, pag. 1: RUPPELDT, F.: Izák Abrahamides – kazateľ, s. 125.



SECVNDA Con
cio.
De Dzahoti a hladu z.
De ohrozności a otupności po.
kuteý i Dzahotey a hladu.
Exordium hic a phum ysaie 9 dabis.
Ktož podruhé následuge mlu
wyti otom: Tzo by Dzahota a
hlad za ohrožnu pokutu byl. A
Hlád obzwlásstě, gestliže gest dluho tr
wagicy: snadno geden každy zmnohych mů
čiti rozymieti mnoże, že gest nálezitě zlé.
Chcežli to čisti a phud Josephum de
Bello Judaico lib: 6. Cap: 11. o obležiny Hie
rusalémské. Aneb a phud Eusebiu
m Ecclica Historia lib: 3. Cap: 6. tzo Eu
sebius ze Josepha wzal. Aneb giných hi
storij o Městech, kteréž dluho trwagiczym
obležením přinuzeni su ku dostosti a
hladu. Nalezně a hrozným a gak stra
šliwé otupným nářzením glad lidi gest
přimuhil. Nedarmo Pisma wé male;

Juuda 331

Obrázok 15 Ukážka rukopisu Abrahamidesovej postily, kde sa súbežne vyskytuje humanistická majuskula, novogotické kreslené písmo, novogotická kurzíva a humanistická kurzíva. Zdroj: LA SNK.

Používanie fraktúry pritom nemuselo byť dané iba lepšou čitateľnosťou a úpravou rukopisu, ale mohlo cielene napodobňovať tlačený text a zaradiť tak rukopis do vyššieho štýlu. Azda nemožno vylúčiť vplyv tlačených českých kázní rozšírených v regióne. Známa bola napr. *Postilla Ewangelitska* od Martina Filadelfa Zámrského (prvé vydanie 1592), z ktorej čerpali aj slovenskí kazatelia.[12] Charakteristika Abrahamidesovej postily zjednodušene ako českej alebo slovakizovanej však nie je dostačujúca. Vo vzťahu k paleografickým špecifikám prameň viac vystihuje jeho zmiešaný jazykový ráz.

Viacjazyčné kázne zachované už od stredoveku sa niekedy považujú za spojenie s hovoreným slovom v minulosti. Väčšinu zo zachovaných textov však nemožno považovať za priame záznamy prednášaných kázní. Mnohé z nich boli určené len na čítanie. V každom prípade však predstavujú množstvo foriem a typov viacjazyčnosti, napr. glosy a preklady, ale aj prepínanie kódu (code-switching), ktoré vo všeobecnosti tvorí najdôležitejšiu tému bádania v tomto smere. Predmetom diskusií zostáva otázka, nakoľko kázne odrážajú prirodzenú bilingválnosť autorov a recipientov a nakoľko by sa mali považovať za produkty rétorického vzdelávania a praxe.[13] V rámci Abrahamidesovej postily popri dominantnej slovakizovanej češtine vystupuje do popredia prepínanie do latinčiny. Podiel latinčiny je natoľko výrazný, že pripomína aktívne a všadeprítomné používanie tohto jazyka v mestskom prostredí. Predpokladom bolo publikum pripravené počúvať texty v latinčine. Ak to tak nebolo, uvažovať možno azda o prekladaní rozsiahlejších pasáží priamo počas čítania kázní. Naopak, v rukopise len sporadicky vystupujú krátke pasáže v gréčtine a veľmi ojedinele aj glosy v hebrejčine. Abrahamides pri citátoch a výrazoch v týchto dvoch jazykoch obvykle zapísal ich preklad a vysvetlenie, čo priamo naznačuje, že neboli všeobecne rozšírené a známe.[14] Prevládajúca čeština okrem hojných slovakizmov zahŕňa aj početné latinizmy, germanizmy a hungarizmy, ktoré v podobe slovníka či registra spracoval v strojopisných poznámkach Fedor Ruppeldt.[15] Prekvapujúco môže pôsobiť, že postila vzhľadom na jazykové pomery v regióne Zvolena a banských miest neobsahuje súvislé zápisy v nemčine ani v maďarčine. Iba okrajovo na ne odkazuje v podobe prísloví a slovných zvratov. Keď napr. apeluje na veriacich, aby sa neprestávali modliť počas moru, hovorí: *Nemczj mluwj: Kreucz lernt betten: Strách se smrti rownáˊ*.[16] Obvykle však argumentuje češtinou a latinčinou, pričom obidva jazyky opakovane prepája s dobovou slovenčinou. Napr. v odpovedi na otázku, či pápežovi vnukol Sv. Duch udeľovanie odpustkov veriacim za návštevu kostolov v Ríme, uvádza: *Na tuto*

12 FIDLEROVÁ, Alena A.: Ke vztahům mezi písařským a tiskařským pravopisným územ v raněnovověkých rukopisech. In: *Bohemica Olomucensia. Filologica Juvenilia*, č. 3, 2009, s. 57 – 58; PAPP, Ingrid: Prítomnosť českej knižnej kultúry v slovenskom písomníctve raného novoveku. In: *Kniha 2021. Zborník o problémoch a dejinách knižnej kultúry*. zost. S. Knapčoková, Martin : Slovenská národná knižnica, 2021, s. 77 – 82; ČIČAJ, Viliam – KEVEHÁZI, Katalin – MONOK, István – VISKOLCZ, Noémi (eds.): *A bányavárosok olvasmányai (Besztercebánya, Körmöcbánya, Selmecbánya) 1533 – 1750. Magyarországi magánkönyvtárak III.* Budapest ; Szeged : OSZK ; Scriptum Rt., 2003, s. 180, č. 83, s. 405, č. 65.

13 ODSTRČILÍK, Jan: Multilingual Medieval Sermons: Sources, Theories and Methods. In: *Medieval Worlds* [online], no. 12, 2020, pp. 140 – 142 [cit. 2022-11-18]. Dostupné na: https://doi.org/10.1553/medievalworlds_no12_2020s140

14 LA SNK, sign. 240 BN 1, pag. 49, 333, 416, 445, 446, 526, 568, 577, 694.

15 LA SNK, sign. 240 AN 3, *Úvahy o kázňovke – rukopise I. Abrahamidesa*, kap. 9: Zvláštnosti reči, 75 s.

16 LA SNK, sign. 240 BN 1, pag. 25.

otazku sprostě, čistě a kratcze po slowensku odpowjdam, že NON.[17] Vzťah k latinčine vyjadrujú predovšetkým citáty antických autorov, ale aj pasáže z Vulgáty. Na druhej strane, k jednostrannému používaniu latinského Písma sa Abrahamides staval kriticky: *... tj, kteřj toliko swau mateřinsku řeč vměgj, a latinského jaziku se nénavčili, obyčegne málo czo wědj o Biblii.*[18] Opäť vystupuje do popredia otázka, či na kázňach zaznievali latinské citáty alebo ich preklad. Postila okrem toho všeobecne odkazuje aj na českú Bibliu. Doterajšie bádanie na základe jazykového rozboru identifikovalo jej konkrétnu redakciu – Kralickú Bibliu vydanú v roku 1593.[19] Môže predstavovať dôležitý zdroj na podrobnejší porovnávací výskum.

OBSAH KÁZNÍ

Vzhľadom na relatívne ucelený stav zachovania tvorí postila Izáka Abrahamidesa obsahom aj rozsahom pestrý súbor rukopisných textov. Jednotlivé texty pôvodne tvorili osobitné zošity. Neskôr boli zviazané v pozmenenom poradí do spoločnej väzby, v záhlaviach však zostalo datovanie podľa jednotlivých nedieľ v rokoch 1600 – 1601. Postilu tvorí tridsať kázní a tri iné útvary. Zväzok sa začína dlhou kázňovou úvahou o more, ktorá rozsahom približne dvojnásobne prevyšuje priemernú dĺžku kázní. O morovej epidémii vypovedá aj ostatných štrnásť kázní s číslovaním I. – XIIII. Ide o tému, ktorej postila venuje najväčšiu pozornosť – zaberá takmer polovicu rukopisu. Mor vysvetľuje ako oprávnený Boží trest za hriešne konanie, ale opisuje aj všedný život počas epidémie. Zároveň ponúka niektoré rady, ako sa uchrániť pred nákazou.[20] Ďalšie štyri kázne Abrahamides zameral na drahotu a hlad. Nedostatok potravín považoval za ešte horší ako mor a rovnako ho vnímal ako Boží trest. Všímal si pritom neúrodu, narastajúce ceny a ich zneužívanie obchodníkmi.[21] Nasledujúce štyri kázne venoval jubilejnému roku 1600, predávaniu odpustkov a učeniu o očistci. Prejavil v nich vo väčšej miere svoju konfesijnú príslušnosť – kritiku pápeža a katolíckej zbožnosti.[22] S témou odpustkov je spojený samostatný krátky text poznámok z kázní nemeckého augustiniánskeho mnícha, „pápeženca" Gottschalka Hollena. Ide o jediný text v postile napísaný iba v latinčine (zakončený česko-latinskou citáciou zdroja).[23] Vzhľadom na použitý jazyk a krátky rozsah je pravdepodobné, že išlo o pomôcku, resp. podklad pre predošlé kázne. Ďalšie tri kázne hovoria o strigách, čarodejniciach a vrahyniach. Abrahamides kritizoval vypočúvanie žien pomocou mučenia a odmietal aj exorcizmus. Pre usvedčené vrahyne však schvaľoval prísne tresty.[24] Špecifický útvar tvorí krátke exordium – úvodná reč venovaná tenkej niti pominuteľného ľudského života, ktorú každému utkal tkáč – Boh.[25] Za ním pokračujú dve kázne zamerané na kritiku úžery. Voľne tým nadväzujú na tému drahoty. Prvá z kázní sa pritom začína neobvykle na rovnakej strane, kde sa končí

17 LA SNK, sign. 240 BN 1, pag. 425; KIŠŠ, Igor: Výskumy Fedora Ruppeldta, s. 12.
18 LA SNK, sign. 240 BN 1, pag. 448.
19 LA SNK, sign. 240 BN 1, pag. 415; RUPPELDT, F.: Izák Abrahamides – kazateľ, s. 126; KIŠŠ, Igor: Etická kritika slovenskej spoločnosti v kázňach Izáka Abrahamidesa z roku 1600 (Rozbor doposiaľ neprebádaného rukopisu Zvolenskej postily). In: *Cirkevné listy*, roč. 134, č. 1 – 2, 2010, s. 37.
20 LA SNK, sign. 240 BN 1, pag. 3 – 306.
21 LA SNK, sign. 240 BN 1, pag. 307 – 409.
22 LA SNK, sign. 240 BN 1, pag. 411 – 520.
23 LA SNK, sign. 240 BN 1, pag. 521 – 522; RUPPELDT, F.: Izák Abrahamides – kazateľ, s. 125 – 126.
24 LA SNK, sign. 240 BN 1, pag. 523 – 588.
25 LA SNK, sign. 240 BN 1, pag. 591 – 595.

text exordia.[26] Nasledujúce dve kázne sú „turecké" (protiosmanské) a reagujú na aktuálne ohrozenie. Prvá kladie dôraz na pokánie ako predpoklad na víťazný boj proti Osmanom a je určená mešťanom. Druhú kazateľ predniesol pred nastúpenými vojakmi. Sústredil sa preto na ich posmelenie, pričom vo väčšej miere využil svetské námety.[27] Postilu napokon uzatvára útvar, ktorý sa datovaním aj lokalizáciou odkláňa od predošlých textov. S prihliadnutím na začiernené miesta v titule ide najpravdepodobnejšie o rozlúčkovú reč (valedikciu) venovanú Jánovi Jacobeiovi († 1612), evanjelickému farárovi v Selciach vo Zvolenskej stolici.[28] Okrem obsahu kázní a rečí tvoria pozoruhodnú zložku postily odkazy na zdroje. Abrahamides sa v marginálnych poznámkach pomerne často odvolával na biblické knihy (najmä starozákonné), ale aj na antických autorov, cirkevných otcov a predstaviteľov reformácie. Odkazy vypovedajú o vzdelanostnej úrovni kazateľa, ktorý využíval široký rozsah zdrojov z teológie, filozofie a histórie.[29] Časovo aj územne blízku obdobu Abrahamidesovho diela predstavovala postila zostavená banskobystrickým kazateľom Jurajom Schmideliom. Obsahovala 44 kázní v slovakizovanej češtine z rokov 1598 – 1607. Dnes nezvestný rukopis zahŕňal aj bližšie neurčenú Abrahamidesovu (podľa iniciálok J. A. H.) kázeň z roku 1600.[30]

TVORBA MODELOV V PLATFORME *TRANSKRIBUS*

Predpoklady na prácu s rukopisom v prostredí *Transkribus* vytvorila digitalizácia Abrahamidesovej postily pracovníkmi Slovenskej národnej knižnice ešte v období okolo roku 2010. Zadná a predná strana väzby a dvojstrany rukopisu boli zdigitalizované v rozlíšení 200 DPI vo formáte JPG. Napriek nezodpovedanej otázke, či bude rozlíšenie dostačujúce, som súbor digitalizátov bez úpravy formátu a kontrastu nahral priamo do softvéru *Transkribus* (verzia 1.17.0) s pôvodným značením naskenovaných snímok 001 – 360. Vzhľadom na vodorovné umiestnenie dvojstrán nebolo potrebné meniť orientáciu digitalizátov. Rukopis pozostáva z blokov textu bez odsekov s pomerne ustálenou hustotou riadkovania (priemerne 25 – 30 riadkov na stranu), pričom jednotlivé kázne oddeľujú vakantné strany. Takáto štruktúra umožnila aplikovať automatickú segmentáciu celej postily. Vygenerované textové rámce s priradeným číslovaním buniek zhora nadol nadobudli štandardne podobu dvoch samostatne značených blokov (v smere čítania zľava doprava) zodpovedajúcich dvom stranám rukopisu. Hranice riadkov v absolútnej väčšine prípadov program rozpoznal správne. Väčšiu mieru nepresnosti preukázal iba v malej časti rukopisu, kde bol použitý atrament s nižšou sýtosťou. Namiesto prevažujúceho hustočierneho išlo o riedky atrament, ktorý nadobudol svetlohnedý odtieň s nízkym kontrastom. Na uvedených miestach automatická segmentácia nezahrnula do rámcov niektoré slová na konci riadkov.[31] Problematické kontúry litier a celých slov som však nedokázal prečítať voľným okom ani

26 LA SNK, sign. 240 BN 1, pag. 595 – 650.

27 LA SNK, sign. 240 BN 1, pag. 655 – 698.

28 LA SNK, sign. 240 BN 1, pag. 699 – 722; SLÁVIK, J.: *Dejiny zvolenského evanjelického...*, s. 241; KIŠŠ, I.: Výskumy Fedora Ruppeldta, s. 13.

29 KIŠŠ, I. Výskumy Fedora Ruppeldta, s. 12; KALAJTZIDIS, J. Superintendent Izák Abrahamides, s. 75 – 77.

30 MOCKO, Ján: Slovenský rukopis z konca XVI. a počiatku XVII. stoletia. Príspevok k dejinám jazyka slovenského. In: *Slovenské pohľady*, roč. XIV, sošit 1, 1894, s. 1 – 5; ČAPLOVIČ, J.: Z osudov starších, s. 41.

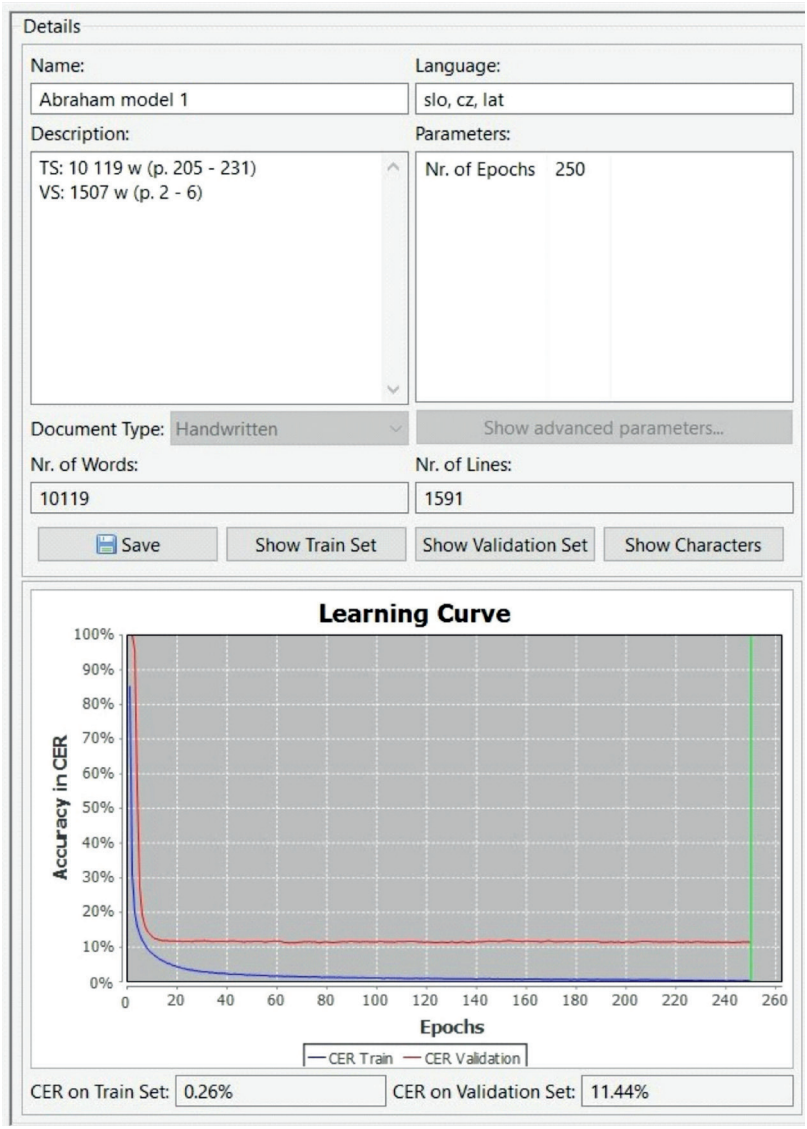
31 LA SNK, sign. 240 BN 1, pag. 358 – 362, 364, 366, 368, 370, 378, 402, 406.

po zväčšení snímky. V tomto kontexte možno použité rozlíšenie digitalizátov považovať za primerané. Korekciu vyžadovala úprava hraníc riadkov a úprava ich číslovania, najmä v prípadoch, ak sa súbežne s blokom textu objavila aj pôvodná marginália alebo kustóda. Bežne sa vyskytlo automatické segmentovanie poznámok, čiar a značiek dopísaných do rukopisu v 20. storočí, rovnako aj automatické segmentovanie paginácie alebo foliácie. Tieto sekundárne prvky som zo segmentácie manuálne odstránil.

Vzhľadom na striedanie rôznych štýlov písma v rukopise bolo na prípravu vzorky strojového učenia so skutočnou hodnotou (Ground Truth) v platforme *Transkribus* potrebné vložiť do upravených riadkov ortograficky verný prepis – transliteráciu. I keď F. Ruppeldt počas výskumu postily realizoval prepis jednej kázne, vyznačuje sa určitými nepresnosťami v čítaní.[32] Spolu so študentmi magisterského štúdia histórie som preto v rámci povinného predmetu základy paleografie realizoval niekoľko sérií prepisov rôznych častí postily. Pre metodiku prepisu predstavovali výzvu osobitosti pisárskej ruky: nie vždy zreteľný rozdiel medzi novogotickou majuskulou a minuskulou, nie vždy zreteľná hranica medzi medzerou a spojením dvoch litier, ako aj posunuté rozlišovacie znaky nad samohláskami. Bežným prípadom je spojenie *Pán Bůh*, pričom dĺžeň nad *a* je posunutý nad *n* a má často podobu bodky. V latinských pasážach v humanistickej kurzíve je zastúpené tradičné skracovanie slov suspenziou a kontrakciou. Tieto skratky som v Transkribe nerozpisoval, nakoľko ich výskyt v prepísaných vzorkách bol obmedzený – nedosahoval odporúčaných 50 znakov potrebných na ich osvojenie. Z rovnakých dôvodov do prepisu neboli zahrnuté texty písané gréčtinou a hebrejčinou. Naopak, rozšírenú skratku vlastného významu v podobe ꝯ (-us) som vkladal pomocou nástroja v Transkribe. V prípade, keď skrátenie obsahovalo literu vyzdvihnutú k vrchnej linke, označil som ju horným indexom. Tento postup sa potom prejavil aj v automatických prepisoch. Spomedzi špecifických znakov s častým výskytom som do prepisu vkladal ligatúru *æ*, ako aj grafémy *ß* a *ÿ*, ktoré si takisto osvojilo strojové učenie. Zachovali sa tým prvky autentickej ortografie.

Prvý model pre automatický prepis tvorila vzorka s počtom 10 119 slov. Jej odporúčané rozdelenie na cvičný súbor a overovací súbor by malo zodpovedať pomeru 10 : 1. Vzorka pracovala s nižším pomerom, kde cvičný súbor tvoril prepis 51 strán. Overovací súbor tvorilo 9 prepísaných strán. Trénovanie modelu v pôvodnom nástroji HTR+ prebiehalo v 250 cykloch. Učenie bolo založené na cvičnom súbore, pričom naučený postup bol automaticky odskúšaný na overovacom súbore. Po skončení trénovania modelu bola chybovosť alfanumerických znakov (CER – character error rate) v cvičnom súbore 0,26 % a v overovacom súbore 11,44 %. Približne po desiatich cykloch sa CER v overovacom súbore už nezlepšovalo (obrázok 16). Presiahnutie 10 % hodnoty CER sa zreteľne prejavilo v nižšej čitateľnosti a zrozumiteľnosti automatického prepisu, kde došlo napr. k zámene litery *h* namiesto správneho *n*, *t* namiesto *č*, *n* namiesto *w*, ale aj menej rušivé *č* namiesto správneho *c*, alebo *i* namiesto *j*. Zámerom pri ďalšom postupe bolo znížiť chybovosť prvého – základného modelu.

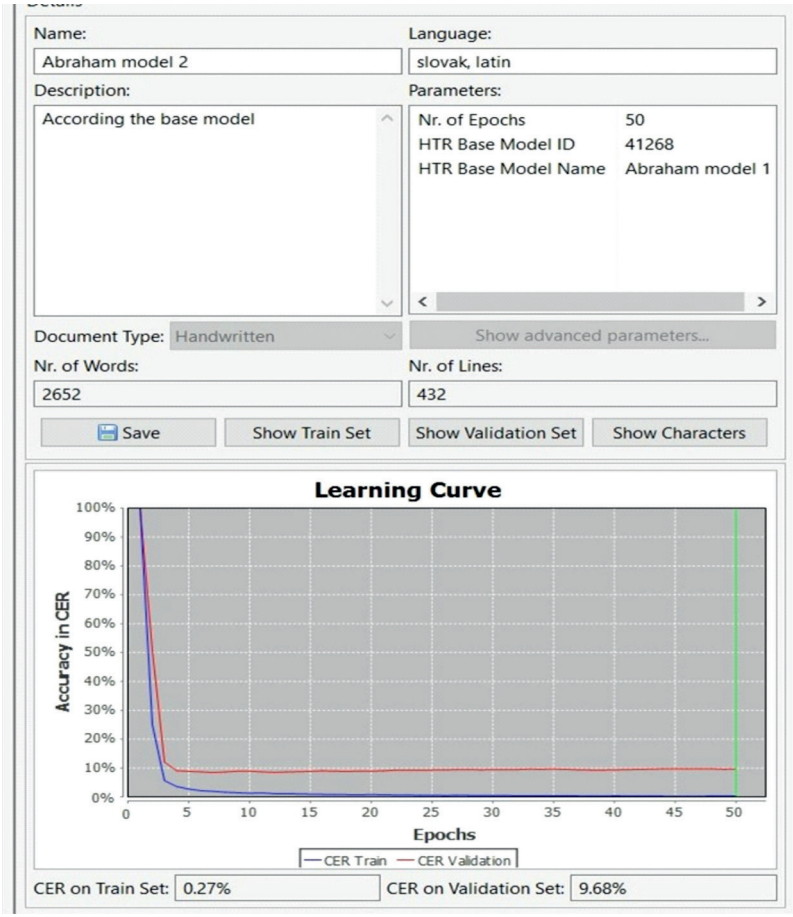
32 LA SNK, sign. 240 AN 3, *Úvahy o kázňovke – rukopise I. Abrahamidesa*, kapitola Prepis kázní, sign. 240 BN 1, pag. 523 – 544.



Obrázok 16 Zobrazenie základného modelu na automatickú transkripciu Abrahamidesovej postily. Zdroj: *Transkribus*.

Druhý model (s použitím základného modelu) pracoval s menšou vzorkou 2 652 slov. Cvičný súbor pozostával zo 16 strán. Overovací súbor tvorili štyri strany. Trénovanie modelu vzhľadom na menší rozsah vzorky prebiehalo v 50 cykloch. Po ich skončení bola chybovosť 0,27 % v cvičnom súbore veľmi podobná prvému modelu. V overovacom súbore mierne poklesla na 9,68 % (obrázok 17). Text automaticky prepísaný v overovacom súbore naďalej obsahoval chyby, viaceré však neboli natoľko rušivé. Napr. obsahoval *l* namiesto správneho *i*, *l* namiesto *t*, *n* namiesto

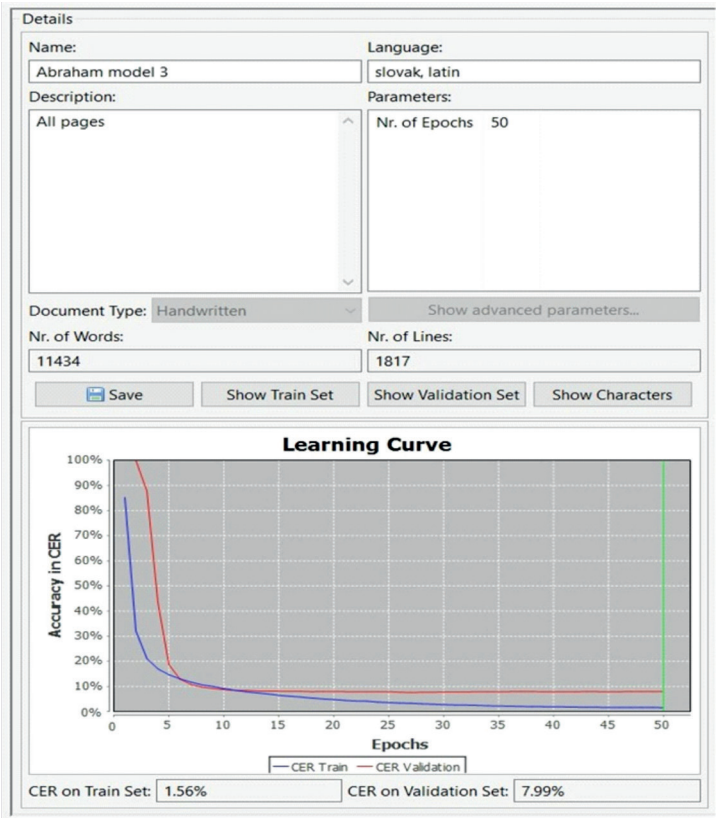
u alebo naopak *u* namiesto *n*. Väčšie množstvo chýb preukazoval prepis v nadpise v humanistickej majuskule aj fraktúre, ktoré zrejme model dostatočne nenatrénoval.



Orázok 17 Zobrazenie druhého modelu s poklesom hodnoty CER v overovacom súbore. Zdroj: *Transkribus*.

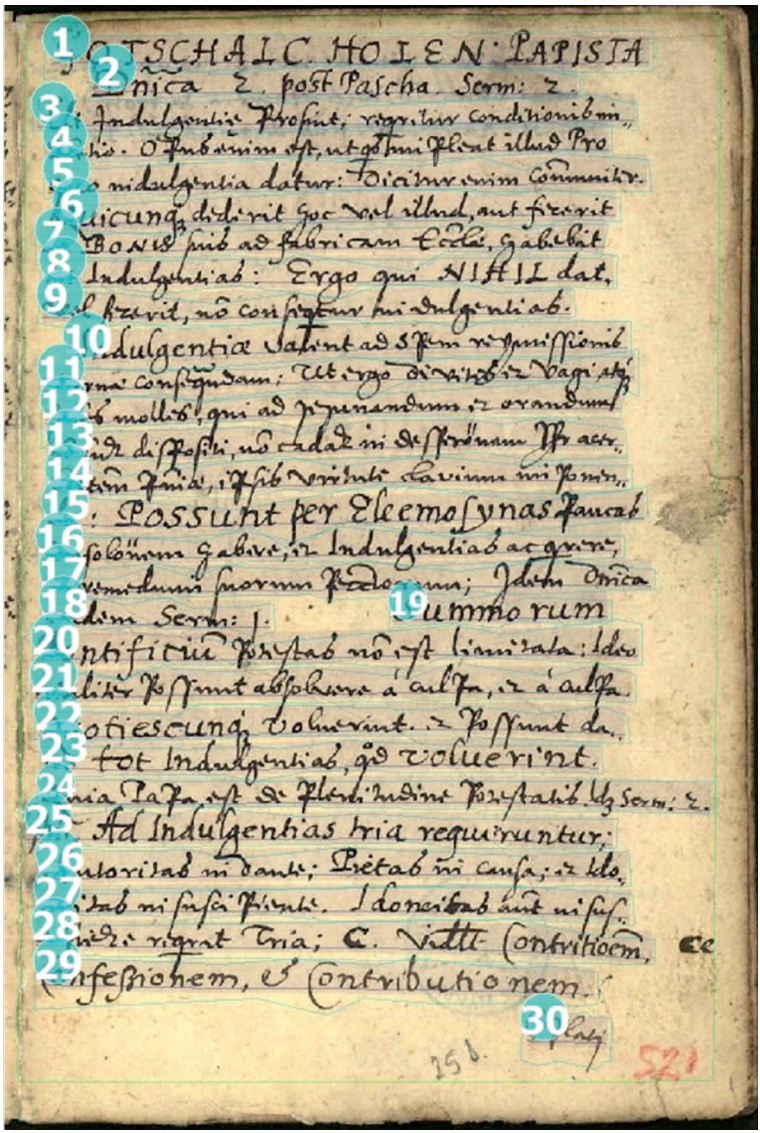
Tretí model pracoval so zlúčením prevažnej väčšiny dovtedy používaných vzoriek v rozsahu 11 434 slov. Cvičný súbor tvorilo 60 strán. Narástol aj rozsah overovacieho súboru s počtom jedenásť strán. Trénovanie prebiehalo v 50 cykloch ako pri predchádzajúcom modeli. Chybovosť v cvičnom súbore predstavovala 1,55 %, avšak v dôležitejšom overovacom súbore klesla hodnota CER na 7,99 % (obrázok 18), čo sa charakterizuje ako použiteľný výsledok. Naďalej sa vyskytovalo zamieňanie *l* namiesto správneho *t*, ktoré zjavne súvisí s málo zreteľným rozdielom medzi obidvomi literami (driek často nie je preložený brvnom). Vyššiu chybovosť naďalej prejavovalo aj čítanie humanistickej majuskuly. Podobne sa v tejto súvislosti prejavili chyby v rímskych číslicach, ale nie v arabských. V rámci diakritiky model často zamieňal *é* a *ě*, avšak tieto rozlišovacie znamienka v neustálenej forme uvádza aj rukopis. Naopak, prepis hlások *ř* a *ů* bol pomerne dôsledný, keďže nie sú zamieňané

s iným druhom diakritiky. Po detailnejšom porovnaní automaticky prepísaného textu s jeho korigovanou verziou sa ukázalo (nástroj Version Comparator), že chybovosť sa často prejavila v slovách, kde chýbal iba dĺžeň, mäkčeň alebo medzera medzi predložkou a slovom. Prepísaný text bol preto vo všeobecnosti zrozumiteľný, vyžadoval však korektúry v rôznom rozsahu závislé od predpokladaného využitia.



Obrázok 18 Zobrazenie tretieho modelu so zlúčenými vzorkami použitými v predchádzajúcich modeloch s ďalším poklesom hodnoty CER v overovacom súbore. Zdroj: *Transkribus*.

Vzhľadom na menší výskyt textov v humanistickom písme počas tvorby modelov vyvstala otázka, do akej miery sa na chybovosti môže podieľať menej zastúpený, a preto menej trénovaný štýl písma. Možnosť pre čiastočnú odpoveď ponúka experiment – automatický prepis pag. 521, ktorá obsahuje iba latinský text (obrázok 19). Napriek predpokladu, že rukopis by mal pozostávať len z humanistického písma, v texte je početne zastúpené aj novogotické písmo. Napr. sloveso *possunt* bolo zapísané najprv humanistickou kurzívou a o niekoľko riadkov nižšie dvakrát už novogotickou kurzívou. Miešanie ortografie mohlo podmieniť zameranie textu v podobe pracovného konceptu – predpokladaných poznámok ku kázňam na rozdiel od ustálenejšieho písomného prejavu v rukopise kázní.



Obrázok 19 Segmentované riadky latinského textu Abrahamidesovej postily. Zdroj: *Transkribus*.

Vytvorenie automatického prepisu bolo založené na základe tretieho modelu s chybovosťou 7,99 %. Po jeho vygenerovaní som realizoval manuálny prepis do segmentovaných riadkov. Po porovnaní obidvoch verzií v Transkribe (obrázok 20) preukázal automatický prepis hodnotu CER až 12,83 %. Zvýšená úroveň chybovosti sa síce prejavila na štatisticky malej vzorke, predsa však možno uviesť niekoľko postrehov. Vyšší podiel chybne prečítaného textu sa opäť objavil v celých slovách zapísaných humanistickou majuskulou, ale aj v majuskulných iniciálach (*U*, *Q*, *I*, *A*,

P), pričom výsledný prepis bol väčšinou nezrozumiteľný. V čítaní kurzívy sa prejavili chyby prítomné už v predchádzajúcich modeloch. Menej rušivé bolo *č* namiesto správneho *c* alebo *g* namiesto *q*, prípadne len posun v proporcii litier: *p* namiesto *P*. Problematickejšie na porozumenie textu bolo čítanie *b* (na konci slov) namiesto správneho *s* alebo *i* v dôsledku posunutej bodky chybne prepísané v podobe *m*, *n*, *r*. Zaujímavý je poznatok, že chybovosť v humanistickej kurzíve sa nejavila vyššia ako v novogotickej. Môže to naznačovať, že tretí model vzhľadom na väčší objem cvičnej vzorky sa už naučil rozpoznávať základné elementy humanistickej kurzívy. Ďalším možným vysvetlením sú menšie ortografické rozdiely medzi obidvomi štýlmi kurzívy (vytvorenými jednou pisárskou rukou) v platforme *Transkribus* na rozdiel od toho, ako sa javia pri manuálnom prepise.

Obrázok 20 Porovnanie automaticky rozpoznaného latinského textu s jeho korigovanou verziou. Zdroj: *Transkribus*.

ZÁVER

Doterajšie skúsenosti s tvorbou prvých modelov ukazujú, že ich efektivita sa zvyšuje s počtom slov zadávaných do prepisovaných vzoriek v kategórii Ground Truth. Pri pokračujúcom výskume bude preto s veľkou pravdepodobnosťou potrebné pracovať s počtom až 15 000 slov, ktoré odporúča ako optimálny rozsah metodika *Transkribus*. Zámerom pri tvorbe ďalších modelov bude znížiť hodnotu CER približne na 5 %, v ideálnom prípade aj nižšiu. S chybovosťou je priamo spojený výskyt latinských skratiek a špecifických grafém. V trénovaných modeloch sa pozitívne prejavilo aplikovanie grafického znamienka pripomínajúceho *y* zakončený oblúčikmi, ktoré v rukopise často uvádza pauzu, rozdiel medzi dvomi vetami. Vo vkladaných vzorkách som mu priradil hodnotu bodky, ktorú v automatickom prepise replikoval aj *Transkribus*. V prípadoch, keď skratky vlastného významu nadobúdajú ustálený fonetický význam, bude po rozšírení trénovanej vzorky (za predpokladu, že sa už skratky vyskytnú 50 alebo viackrát) účelné ich rozpisovanie. Môže ísť napr. o skratky *pro-* alebo *per-* v rukopise značené literou *p* so šikmo alebo vodorovne prekríženou dolnou dĺžkou. Tieto rozpísané skratky by sa už potom nemali podieľať na chybovosti a zároveň prispejú k zvyšovaniu čitateľnosti automatického prepisu. Súčasne tým vznikne posun od prvotnej transliterácie k transkripcii, ktorá poskytuje väčší priestor na využitie textového výstupu na študijné účely v rôznych exportovaných formátoch (napr. docx) alebo súbežne so snímkami digitalizovaného rukopisu v prostredí *Transkribus Lite*.[33] Po nevyhnutnej korektúre automaticky segmentovaných textových rámcov a riadkov v celom súbore bude možné na základe modelu s nízkou chybovosťou pristúpiť ku komplexnej automatickej transkripcii. Môže tak vzniknúť editovateľný „živý" text, ktorý umožňuje štandardnú i rozšírenú edičnú úpravu. Automatický prepis sprístupní paleograficky náročnejší, avšak jazykovo a obsahovo atraktívny prameň, ktorý navyše odráža kultúrne špecifiká slovenského prostredia v stredoeurópskom rámci. Vytvorenie funkčného modelu na rozpoznávanie rukopisu Izáka Abrahamidesa môže v neposlednom rade poskytnúť východisko na štúdium ďalších rukopisov spojených s touto historickou postavou v digitálnom prostredí.

33 NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 63 – 65.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

BUCHWALD, Georg: Beiträge zur Kenntniss der evangelischen Geistlichen und Lehrer Oesterreichs aus den Wittenberger Ordiniertenbüchern seit dem Jahre 1573. In: *Jahrbuch der Gesellschaft für die Geschichte des Protestantismus in Oesterreich* [online]. hrsg. Georg Loesche, 21. Jahrgang, 1900, s. 113 – 128 [cit. 2022-11-21]. Dostupné: https://anno.onb.ac.at/cgi-content/anno-plus?aid=jgp&datum=1900&page=117&size=45

CSEPREGI, Zoltán: *Evangélikus lelkészek Magyarországon (ELEM) II/1. A zsolnai zsinattól (1610) a soproni országgyűlésig (1681). II/1: Nyugat-Magyarország (a dunántúli, a bajmóci* és a *felső-dunamelléki egyházkerület)* [online]. Budapest : RECITI, 2020 [cit. 2022-11-21]. Dostupné na: https://mek.oszk.hu/23000/23048/23048.pdf

ČAPLOVIČ, Ján: Z osudov starších slovenských súkromných knižníc. In: *Knižnica : časopis pre knižnú kultúru*, roč. III – IV, č. 2, 1951, s. 25 – 42.

ČIČAJ, Viliam – KEVEHÁZI, Katalin – MONOK, István – VISKOLCZ, Noémi (eds.): *A bányavárosok olvasmányai (Besztercebánya, Körmöcbánya, Selmecbánya) 1533 – 1750. Magyarországi magánkönyvtárak III.* Budapest ; Szeged : OSZK ; Scriptum Rt., 2003.

FIDLEROVÁ, Alena A.: Ke vztahům mezi písařským *a* tiskařským pravopisným územ v raněnovověkých rukopisech. In: *Bohemica Olomucensia. Filologica Juvenilia*, č. 3, 2009, s. 52 – 59.

HOLUBY, Jozef Ľudovít: Životopis superintendenta Mr. Izáka Abrahamidesa. Z pôvodného rukopisu Laučekovho, r. 1795 písaného, a v Zay-Uhrovskom archíve opatrovaného. In: *Cirkevné listy*, roč. XXIV, č. 4, 1910, s. 114 – 117.

HOLUBY, Jozef Ľudovít: Životopis superintendenta Mr. Izáka Abrahamidesa. Z pôvodného rukopisu Laučekovho, r. 1795 písaného, a v Zay-Uhrovskom archíve opatrovaného. In: *Cirkevné listy*, roč. XXIV, č. 6, 1910, s. 171 – 175.

KALAJTZIDIS, Ján: Superintendent Izák Abrahamides Hrochotský a jeho Zvolenská postila. In: *Historia Ecclesiastica*, roč. XI, č. 2, 2020, s. 68 – 78.

KIŠŠ, Igor: Etická kritika slovenskej spoločnosti v kázňach Izáka Abrahamidesa z roku 1600 (Rozbor doposiaľ neprebádaného rukopisu Zvolenskej postily). In: *Cirkevné listy*, roč. 134, č. 1 – 2, 2010, s. 36 – 43.

KIŠŠ, Igor: Výskumy Fedora Ruppeldta na Abrahamidesovej postile z r. 1600 – 1601. In: *Cirkevné listy*, roč. 134, č. 5, 2010, s. 8 – 14.

KIŠŠ, Igor: Úsilie Izáka Abrahamidesa o výchovu k humánnej spoločnosti. In: *Slovenské pohľady*, roč. IV. + 129, č. 5, 2013, s. 61 – 74.

KOVAČKA, Miloš et al.: Osobnosti Žilinskej synody. Malý biobibliografický slovník. In: *Akty a závery – zákony a ustanovenia Žilinskej synody.* ed. M. Kovačka, Martin: Slovenská národná knižnica, 2010, s. 66 – 89.

MOCKO, Ján: Slovenský rukopis z konca XVI. a počiatku XVII. storočia. Príspevok k dejinám

jazyka slovenského. In: *Slovenské pohľady*, roč. XIV, sošit 1, 1894, s. 1 – 10.

NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 53 – 67.

ODSTRČILÍK, Jan: Multilingual Medieval Sermons: Sources, Theories and Methods. In: *Medieval Worlds* [online], no. 12, 2020, pp. 140-147 [cit. 2022-11-21]. Dostupné na: https://doi.org/10.1553/medievalworlds_no12_2020s140

Oratio Exequialis, Illustrissimo p. m. Comiti ac Domino, D. Georgio Thvrzoni de Bethlemfalwa [online]. Dicta et Recitata ab ISAACO Abrahamide HROCHOTIO, Levtschoviae : Daniel Schultz, 1617 [cit. 2022-11-18]. Dostupné na: http://147.213.131.4:85/digi/mf/Mf_084/start.htm

ORMIS, Ján V. (ed.): Listy Sama Chalupku. In: *Litteraria. Štúdie a dokumenty II.* red. O. Čepan, Bratislava : Vydavateľstvo Slovenskej akadémie vied, 1959, s. 89 – 185.

PAPP, Ingrid: Prítomnosť českej knižnej kultúry v slovenskom písomníctve raného novoveku. In: *Kniha 2021. Zborník o problémoch a dejinách knižnej kultúry.* zost. S. Knapčoková. zost. S. Knapčová, Martin : Slovenská národná knižnica, 2021, s. 74 – 85.

PONGRÁCZ, Denis: *Atlas osobných pečatí I.* Bratislava : Vydal JUDr. Mikuláš Trstenský vlastným nákladom, 2019.

RUPPELDT, Fedor: Izák Abrahamides – kazateľ. In: *Cirkevné listy*, roč. 76, č. 7 – 8, 1963, s. 124 – 127.

SLÁVIK, Ján: Vysvätenie Izáka Abrahamidesa Hrochoťského na biskupstvo. In: *Cirkevné listy*, roč. XXXIII, č. 4 – 5, 1919, s. 104 – 108.

SLÁVIK, Ján: *Dejiny zvolenského evanjelického a. v. bratstva a seniorátu.* Banská Štiavnica : Tlačou a nákladom vdovy a syna Augusta Joergesa, 1921.

Slovenská národná knižnica v Martine – Literárny archív, Zbierky a fondy, Literárne rukopisy, XXXI, 240 Fedor Ruppeldt.

ŽILÁK, Ondrej: Izák Abrahamides. In: *Cirkevné listy*, roč. 84, č. 9, 1971, s. 136 – 140.

KAPITOLA 3

SPRÍSTUPNENIE CSÁKÓSOVHO KATALÓGU KOREŠPONDENCIE KOHÁRYOVCOV POMOCOU AUTOMATICKEJ TRANSKRIPCIE

Imrich Nagy
Univerzita Mateja Bela v Banskej Bystrici; Filozofická fakulta; Katedra histórie
E-mail: imrich.nagy@umb.sk

ABSTRAKT

Dobové archívne pomôcky – registre k nespracovaným fondom sú vhodným objektom na digitalizáciu metódou automatickej transkripcie. Príkladom je fond korešpondencie významného uhorského šľachtického rodu Kohárovcov deponovaný v Štátnom archíve v Banskej Bystrici, ku ktorému v polovici 20. storočia spracoval číselný katalóg s podrobnými regestami listov bratislavský mestský archivár Jozef Ján Csákós. Príprava modelu na jeho automatickú transkripciu v platforme *Transkribus* odhalila problém nemožnosti aplikovať automatickú segmentáciu textu. Dôvodom je forma rukopisných tabuliek, do ktorých Csákós zapísal svoj katalóg. Pri takomto type textu je nevyhnutné manuálne segmentovať jednotlivé bloky textu v tabuľke a stanoviť poradie ich čítania. Postupným pridaním manuálne prepísaných a korigovaných strán (spolu 191) do vzorky Ground Truth sa podarilo vyvinúť úspešný model na automatickú transkripciu s chybovosťou okolo 2 % na overovacom súbore. Realizácia digitalizácie Csákósovho katalógu korešpondencie Kohárovcov prináša merateľnú pozitívnu hodnotu za vynaložené prostriedky a ponúka možnosť ďalšieho využitia pri digitalizácii jednotlivých listov z korešpondencie, resp. iných rukopisných materiálov z pozostalosti J. Csákósa.

Kľúčové slová: Kohárovci, katalóg, korešpondencia, Jozef Ján Csákós, automatická transkripcia, *Transkribus*

ABSTRACT

Access to the Csákós catalog of correspondence of the Koháry family using automatic transcription

Historical archival aids – registers of unprocessed funds are a suitable object for digitization using the automatic transcription method. An example is the Fund of Correspondence of the important Hungarian noble family Koháry deposited in the State Archives in Banská Bystrica, for which in the middle of the 20th century a numerical catalog with detailed registers of letters was compiled by Bratislava city archivist Jozef Ján Csákós. The preparation of the model for its automatic transcription in the *Transkribus* platform encountered the problem of the inability of applying automatic text segmentation. The

reason is the form of manuscript tables in which Csákós wrote his catalog. With this type of text, it is necessary to manually segment individual blocks of text in the table and determine the order of their reading. By gradually adding manually transcribed and corrected pages (191 in total) to the Ground Truth sample, it was possible to develop a successful automatic transcription model with an error rate of around 2% on validation set. The realization of the digitization of Csákós's catalog of the correspondence of the Koháry family brings a measurable positive value for the funds spent and offers the possibility for further use in the digitization of individual letters from the correspondence, respectively other manuscript materials from the estate of J. Csákós.

Keywords: the Koháry family, catalog, correspondence, Jozef Ján Csákós, automatic transcription, *Transkribus*

ÚVOD

Za hlavnú cieľovú skupinu na aplikáciu metódy automatickej transkripcie historických rukopisných textov možno označiť pamäťové inštitúcie – archívy, ktoré by mali zo svojej podstaty okrem iného ochraňovať a sprístupňovať archívne dokumenty.[1] Práve tieto dva body zo zákonnej definície poslania archívu môže totiž digitalizácia archívneho dokumentu formou jej automatickej transkripcie pomôcť naplniť. Súčasný stav vývoja technológií však neumožňuje masové a strojové nasadenie tejto metódy, čo vyvoláva na strane archívov minimálne zdržanlivosť a otázky o jej účelnosti a praktickej využiteľnosti. Je to otázka získanej hodnoty za vynaložené prostriedky (rozumiem pod tým nielen peniaze, ale aj, či predovšetkým prácu odborného archívneho pracovníka). Ide najmä o potrebu časovo i odborne náročného postupu vytvárania modelov automatickej transkripcie limitovaných pre jednotlivé dokumenty s konkrétnym rukopisom, čo zvyčajne predstavuje len obmedzený počet strán. Na tomto mieste je vhodné pripomenúť, že so zdokonaľovaním technológie, ktorého predpokladom je práve vytváranie nových modelov automatickej transkripcie, sa otvára cesta aj na tvorbu univerzálnych modelov schopných automatickej transkripcie typovo (napr. formálne, chronologicky, obsahovo atď.) analogických historických dokumentov. Ak aj nebudem zdôrazňovať, že bez nášho vlastného, územne, jazykovo či provenienčne slovacikálneho vkladu do zdokonaľovania tejto progresívnej technológie prostredníctvom tvorby modelov jej praktickú využiteľnosť a aplikovateľnosť v našich podmienkach iba zbytočne oddialime, musím uviesť, že z hľadiska každodenných potrieb archívov už teraz existuje oblasť, ktorá si vyslovene žiada jej nasadenie. Na mysli mám digitalizáciu dobových archívnych pomôcok, inventárov, súpisov a katalógov, ktoré neraz stále plnia svoju pôvodnú úlohu základnej informácie o príslušnom archívnom fonde a orientácie v ňom a umožňujú cielený prístup k jednotlivým dokumentom, ktorý by inak nebol možný. Účelnosť digitalizácie dobových archívnych pomôcok umocňuje aj skutočnosť, že zvyčajne ide o stranovo rozsiahle texty písané jedným autorským rukopisom v presne formátovanej, unifikovanej podobe, čo spĺňa základné predpoklady na aplikovanie nástroja automatickej transkripcie. To, že takto získaná hodnota za vynaložené prostriedky je naozaj relevantná, preukážem na príklade takejto dobovej

1 § 2 zákona č. 395/2002 Z. z. o archívoch a registratúrach a o doplnení niektorých zákonov v znení neskorších predpisov. Dostupné tiež na: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2002/395/

archívnej pomôcky, ktorou je rukopisný katalóg ku korešpondencii Kohárovcov z fondov Štátneho archívu v Banskej Bystrici.

KOHÁRYOVCI A ICH RODOVÝ ARCHÍV

Šľachtický rod Kohárovcov je príkladom úspešného príbehu rodiny, ktorá sa vlastným pričinením prepracovala medzi elitu štátu. Pôvodne patrili medzi nižšiu šľachtu – ako príklad možno uviesť Imricha Kohárho, ktorý bol v druhej polovici 16. storočia kapitánom mesta Krupina a veliteľom hradu Čabraď.[2] Vďaka svojim verným službám vládnucim Habsburgovcom, ktoré v priebehu 16. a 17. storočia vykonali najmä na poli protiosmanských bojov – veď Štefan Kohári I. v nich dokonca obetoval aj vlastný život, keď v roku 1664 padol v bitke pri Leviciach[3] – príslušníci rodiny získali povýšenie medzi šľachtickú elitu štátu, vyjadrené udelením barónskeho titulu, nadobudnutím rozsiahlych majetkov a získaním vysokých stoličných a krajinských funkcií. Postupne sa stali jedným z najbohatších a najmocnejších šľachtických rodov, pričom významná časť ich majetkov sa nachádzala na území dnešného Slovenska – v historickej Novohradskej a Gemerskej stolici. Svoj vrchol dosiahla rodina v roku 1815, keď panovník František I. Rakúsky (cisár František II. Habsburský) udelil Františkovi Kohárymu kniežací titul. S Františkom kohárovský rod po meči vymiera. Stihol však zaistiť spojenie svojho rodu s mimoriadne vplyvným rodom Coburgovcov, ktorý bol súčasťou kniežacieho a kráľovského rodu sasko-kobursko-gothskej dynastie, sobášom svojej dcéry Márie Antónie Kohárovej a Ferdinanda Juraja Coburga. Coburgovci sa tak stali dedičmi rozsiahlych kohárovských majetkov aj na území Slovenska. Ich centrom a rodovým sídlom bol vtedy kaštieľ vo Svätom Antone, kde sa uchovávali aj rodinné dokumenty a všetky písomnosti týkajúce sa majetkov rodiny. Od 18. storočia rodina angažovala už aj osobitných archivárov, ktorí mali na starosti predovšetkým triedenie rodinných dokumentov a vyhotovenie prvých archívnych pomôcok na základnú orientáciu v mimoriadne rozsiahlom rodovom archíve. Fond bol rozdelený do piatich častí označených rímskymi číslami I. – V. Doboví archivári pri usporadúvaní dokumentov nedodržali chronologický ani vecný systém zoradenia. Po roku 1918, keď sa správcom kohárovsko-coburgovských majetkov stal Štátny pozemkový úrad, bol rodinný archív prevezený zo Sv. Antona do Bratislavy. Po 2. svetovej vojne v súvislosti s konfiškáciami majetkov šľachty a s pozemkovou reformou zriadilo Povereníctvo poľnohospodárstva a pozemkovej reformy v roku 1947 Poľnohospodársky archív pri Povereníctve poľnohospodárstva, ktorého úlohou bolo prevziať do vlastníctva štátu archívy skonfiškovaných veľkostatkov a lesných závodov.[4] Poľnohospodársky archív teda prevzal aj rodový archív Kohári-Coburgovcov a deponoval ho vo svojej banskobystrickej pobočke. Dnes je súčasťou fondov Štátneho archívu v Banskej Bystrici,[5] ale dodnes je nespracovaný, t. j. nemá vypracovanú archívnu pomôcku.

2 MATUNÁK, Michal: *Krupinskí hradní kapitáni*. ed. Miroslav Lukáč, Krupina : Kultúrne centrum a Múzeum Andreja Sládkoviča v Krupine v spolupráci s Mestom Krupina, 2011, s. 14.

3 KOPČAN, Vojtech: *Turecké nebezpečenstvo a Slovensko*. Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 1986, s. 141.

4 CHALUPECKÝ, Ivan: K problematike štátnych a cirkevných archívov na Slovensku. In: *Z minulosti Spiša. Ročenka Spišského dejepisného spolku v Levoči – XXIV, roč. 2016*. Levoča : Spišský dejepisný spolok Levoča, 2016, s. 188.

5 Štátny archív Banská Bystrica, fond Kohári – Coburg (1241) 1321 – 1945.

Väčšina dokumentov kohárYovského rodového archívu bola v 70. a 80. rokov 20. storočia zosnímaná na mikrofilmy pre potreby vtedajšieho Maďarského krajinského archívu (MOL). Dnes je k dispozícii pre bádateľov v Maďarskom národnom archíve v zbierke českých a slovenských mikrofilmov (MNL-OL C). Ani radenie týchto mikrofilmov však nijako nezohľadňuje formu (listy, listiny, účtovné záznamy) a obsah (usporiadanie z hľadiska chronologického či tematického) dokumentov, čo mimoriadne sťažuje bádanie v nich. Možno teda konštatovať, že nespracovanosť fondu v jeho listinnej, resp. zosnímanej podobe na mikrofilmy je dôvodom, prečo ešte v historiografii nedisponujeme ucelenou monografiou o KohárYovcoch.

KOREŠPONDENCIA KOHÁRYOVCOV

Listy z korešpondencie KohárYovcov sú zaradené do I., IV. a V. časti rodového archívu. V I. časti sú to najmä listy adresované Štefanovi Kohárymu I. a II. Obsahovo sú zaujímavé najmä tie, ktoré sa týkajú protiosmanských bojov, či protihabsburských povstaní Štefana Bočkaja a Imricha Tököliho. Sú tu však aj listy, ktoré si vymieňali navzájom členovia rodiny a týkali sa správy majetkov. V časti IV. rodového archívu sa nachádzajú listy súkromného a hospodársko-správneho charakteru, ale aj listy k záležitostiam na celoštátnej úrovni. Napísané boli v rokoch 1524 – 1825. Listy s hospodárskym obsahom, ktoré zvyčajne písali šafári, správcovia majetkov a prefekti svojim pánom, hovoria o hospodárení na majetkoch, investíciách, stavebných projektoch a tiež o situácii poddaných. Celouhorská problematika sa objavuje najmä v listoch vtedajších významných politikov (napr. Jána a Mikuláša Pálfiho, Žigmunda Esterházyho, Štefana Zičiho, princa Eugena Savojského a ďalších). Obsahujú najmä informácie o vojenských aktivitách, následkoch bojov s Osmanmi, ale aj o zahraničnopolitických udalostiach najmä vojnového charakteru. Podobný charakter majú aj listy z V. časti rodového archívu, ktoré sú datované v rokoch 1543 – 1849. Aj po obsahovej stránke tieto listy nadväzujú na predchádzajúcu časť. Zaujímavé sú správy o osmanskom plienení a výzvach Osmanov obyvateľom z poplatných území. Nachádzajú sa tu aj listy dvorskej vojnovej rady o prípravách na vojenské akcie proti Osmanom či o stavovských povstaniach. Objavujú sa aj správy o hladomore a epidémiách, ktoré boli prirodzenými sprievodnými javmi vojenských konfliktov. Napokon sú tu aj listy adresované KohárYovcom ako hlavným županom Hontianskej stolice. Korešpondencia KohárYovcov je veľmi cenným historickým prameňom, ktorý ponúka možnosť nahliadnuť do dejín každodennosti, hospodárskych dejín, vojenských dejín, dejín šľachty i dosahov makrohistórie na mikrohistóriu. Hoci možnosti využitia tohto prameňa historikmi sú, ako sme už uviedli, limitované, neprestáva priťahovať ich pozornosť. Ako príklad tu možno uviesť práce maďarského historika Zoltána Komjátiho[6] či slovenskej historičky Tünde Lengyelovej.[7]

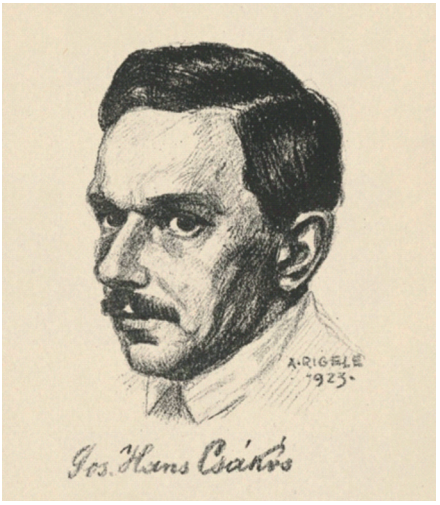
6 KOMJÁTI, Zoltán Igor: „... Az mit hallottam, kötelességem szerint akarám Nagyságodnak tudtára adnom...” Adalékok a híráramlás és a hírhálózat történetéhez Felső-Magyarországon KohárY István füleki főkapitány levelezésének tükrében (1672 – 1682). In: *Fons (Forráskutatás és Történeti Segédtudományok)*, roč. XVII, č. 1, 2010, s. 113 – 140.

7 LENGYELOVÁ, Tünde: „Ja potrebujem peniaze viac, ako tvoja milosť“: korešpondencia Judity Balassovej so synom Štefanom II. Kohárym. In: In: *„Za Boha, kráľa a vlasť!“: KohárYovci v uhorských dejinách. Zborník príspevkov z odborného seminára, ktorý sa konal 23. – 24. septembra 2015 v Múzeu vo Svätom Antone*. ed. M. Ďurianová, Zvolen : Národné lesnícke centrum, 2016, s. 9 – 19.

DOBOVÁ ARCHÍVNA POMÔCKA – CSÁKÓSOV KATALÓG KU KOREŠPONDENCII KOHÁRYOVCOV

Korešpondencia Kohároyovcov bola usporiadaná do fascikulov podľa adresátov v prvej polovici 19. storočia. Neprihliadalo sa pri tom na ich obsah, takže sa pomiešané vedľa seba ocitli listy hospodárskeho, osobného aj administratívneho charakteru.[8] Eugen Neuenschwander, ktorý od začiatku 20. storočia zastával funkciu archivára rodového archívu, usporiadal listy zo 17. – 18. storočia v rámci fascikulov do chronologického poradia a začal s prácami na číselnom katalógu, v ktorom ku každému listu pripojil aj krátky regest. Katalóg v dvoch zväzkoch obsahuje položky od 1 do 2000, ktoré označujú jednotlivé listy.

Tu nadviazal na Neuenschwanderovu prácu v 30. rokoch 20. storočia Jozef Ján Csákós (1883 – 1957). Mladý Jozef Ján sa u svojho nevlastného otca Michala Csákósa vyučil ryteckému remeslu. Rytecká prax mala pravdepodobne vplyv na jeho charakteristický krasopisný úhľadný štýl, ale zároveň dosť ťažko čitateľný rukopis, ktorý sa zachoval v jeho autorských rukopisných dokumentoch.[9] Do roku 1922 pracoval J. Csákós v mestskom múzeu v Bratislave ako kustód, pričom sa venoval najmä rukopisným zbierkam. Pre zhoršujúci sa zdravotný stav musel požiadať o uvoľnenie z pracovnej pozície. Do roku 1945 potom pracoval ako mestský archivár.[10] Do tohto obdobia patrí aj jeho práca na katalógu korešpondencie Kohároyovcov.



Obrázok 21 Rigele, Alojz: Portrét Jozefa Jána Csákósa. Litografia. 1923. Galéria mesta Bratislavy. Zdroj: https://www.webumenia.sk/dielo/SVK:GMB.C_13385

8 GOMBOS, János: A Koháry-Coburg család besztercebányai családi levéltára. In: *Levéltári Szemle,* roč. 43, č. 2, 1997, s. 37.

9 FRANCOVÁ, Zuzana: Jozef Ján Csákós (1883 – 1957), bývalý kurátor múzea. In: *Zborník Múzea mesta Bratislavy*, roč. 29. Bratislava : Múzeum mesta Bratislavy, 2017, s. 240.

10 FONÓD, Zoltán (ed.): *A cseh/szlovákiai magyar irodalom lexikona 1918 – 2014.* 2. oprav. vyd. Bratislava : Madách-Posonium, 2004, s. 61.

Csákós pokračoval v Neuenschwanderovom číselnom katalógu a v rokoch 1944 – 1945 spísal položky od čísla 2001 až po číslo 8633, ktoré zodpovedali jednotlivým listom.[11] Doplnil ich obsiahlymi regestami, v niektorých prípadoch možno hovoriť skôr o prepisoch pôvodných listov. Spomenúť treba však aj to, že ešte v rokoch 1936 – 1937 vypracoval stručný číselný súpis aj ku korešpondencii z V. časti rodového archívu, ktorý obsahoval položky 1 – 19758.[12] Predmetom nášho výskumu sa však stal Csákósov podrobný katalóg ku korešpondencii nachádzajúcej sa v časti IV rodového archívu. Tento katalóg obsahuje regesty 6632 listov rôzneho rozsahu (v niektorých prípadoch obsahujúce aj výťahy, resp. preklady originálnych listov in extenso) väčšinou v maďarskom, ale čiastočne aj v nemeckom a latinskom jazyku spísané v rukopise moderným kurzívnym písmom v podobe tabuliek na 4140 stranách formátu A3 (250 x 400 mm) v zošitovej väzbe. Jednotlivé zošity pozostávajú priemerne z 10 listov, t. j. 20 strán. Všetky strany sú priebežne paginované v dvoch korpusoch – prvý korpus od 1 po 3000 a druhý od 1 po 1140.

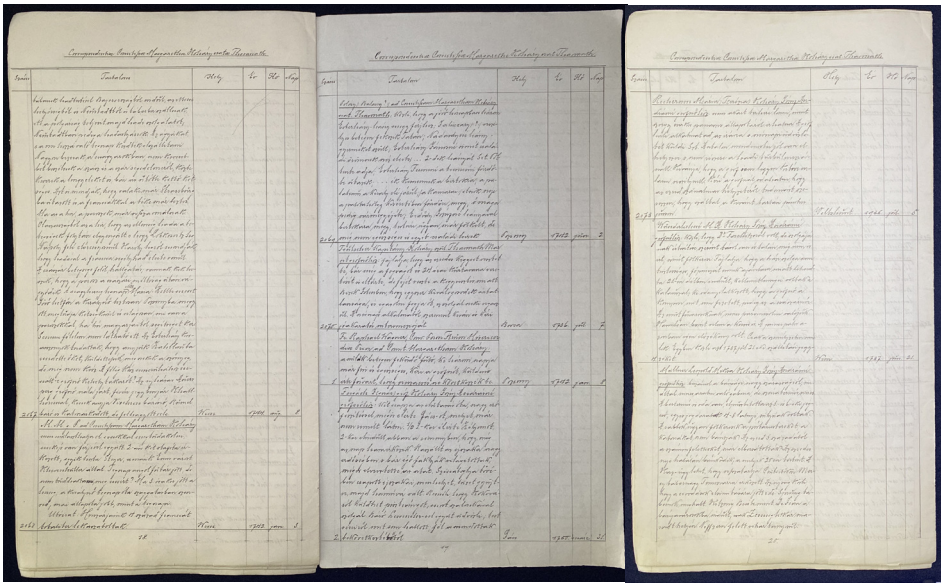
NASNÍMANIE DOKUMENTU A JEHO SEGMENTÁCIA NA AUTOMATICKÚ TRANSKRIPCIU

Prvým krokom k digitalizácii Csákósovho katalógu prostredníctvom platformy *Transkribus* na automatickú transkripciu rukopisných textov bolo vyhotovenie digitalizátov. Dvojstrany katalógu som nasnímal na digitálny fotoaparát mobilného telefónu iPhone 11 Pro pomocou jednoduchého statívu so zabudovaným neutrálnym osvetlením v podobe tzv. ScanTentu, ktorý vyvinuli riešitelia projektu READ.[13] Týmto postupom som vyhotovil 2075 snímok s rozmermi 4032 x 3024 pixelov pri rozlíšení 192 DPI zachytávajúcich dvojstránku originálneho dokumentu. Optika fotografickej sústavy použitého mobilného telefónu nedokázala vyvážiť kontrast medzi samostatnou stranou a tmavým podkladom ScanTentu pri snímke prvej a poslednej strany zošita. Snímky boli z tohto dôvodu preexponované, čo sa negatívne premietlo do výsledku rozpoznávania rukopisného textu. Vzhľadom na veľký počet takýchto snímok (každá 10. snímka) som dodatočne pristúpil k opätovnému nasnímaniu problematických strán tak, že som do jedného záberu spojil poslednú a prvú stranu nadväzujúcich zošitov.

11 OTRUBA, Štefan: *Štátny archív v Banskej Bystrici: sprievodca po archívnych fondoch II.* Bratislava : Slovenská archívna správa, 1969, s. 17.

12 GOMBOS, J.: A Koháry-Coburg család besztercebányai családi levéltára, s. 38.

13 The ScanTent: Professional scanning with your smartphone. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-12-10]. Dostupné na: https://readcoop.eu/scantent/



Obrázok 22 Rozdiel v kvalite snímky dvojstrany a samostatnej strany originálu dokumentu. Zdroj: Fotoarchív autora.

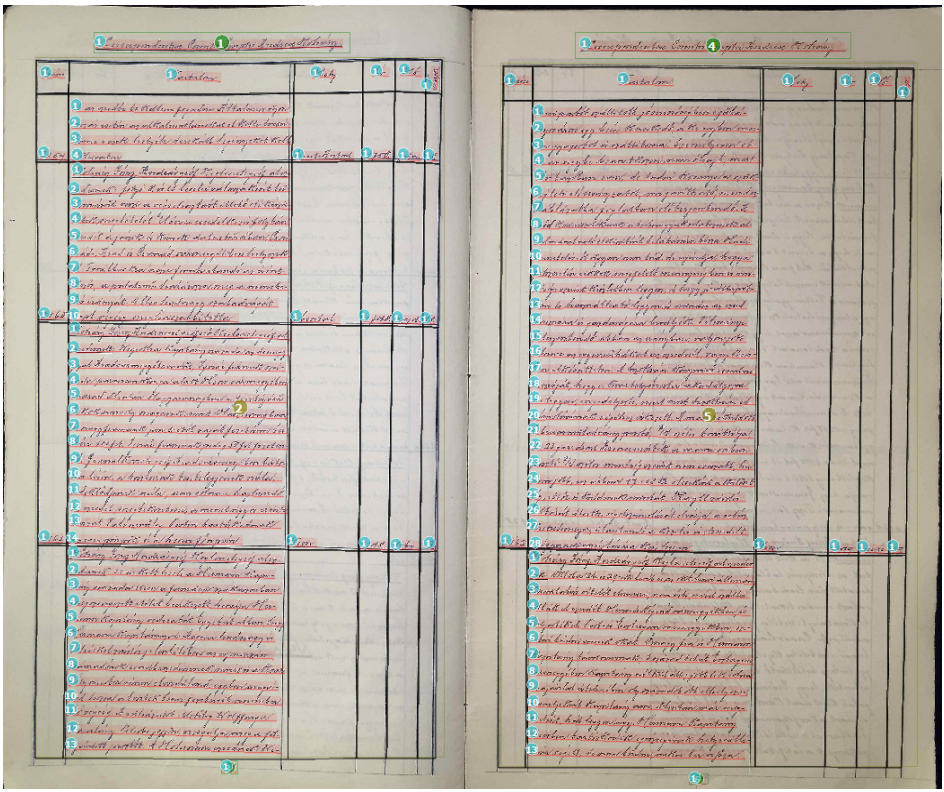
Všetky snímky Csákósovho katalógu som nakoniec importoval pomocou *Transkribus Expert Client* do platformy *Transkribus*. Tento užívateľský program je zo strany vývojárov platformy *Transkribus* vzorovo aktualizovaný – momentálne je už dostupná jeho verzia s označením v1.22.1.[14] Platforma však ponúka aj verziu pre webové prehliadače *Transkribus Lite* (aktuálne vo verzii v2.2.2.3), ktorá takisto umožňuje využiť základné funkcionality platformy, počnúc importovaním digitalizátov, cez trénovanie modelu až po rozpoznávanie rukopisného textu a úpravu kompletných zbierok.[15]

Ďalším krokom po importovaní digitalizátu je jeho segmentácia, 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 používateľa potom zostáva len kontrola a oprava (napr. spresnenie hranice riadku, zmena poradia čítania a pod.). Csákós však svoj katalóg vpísal do tabuľky, ktorú narysoval na každú stranu. Hoci autori platformy neustále pracujú na zdokonaľovaní automatizovaných procesov, v čase prípravy modelu pre Csákósov katalóg ešte modul na automatické rozpoznanie blokov textu v tabuľke nebol pre tento rukopis aplikovateľný. Preto som vykonal manuálne vymedzenie týchto blokov, ktoré som 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, čo som upravil na čítanie po riadkoch). Takto vytvorené rámce umožňuje

14 Download Transkribus. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-12-20]. Dostupné na: https://readcoop.eu/transkribus/download/

15 *Transkribus Lite: Automatically transcribe, comfortably edit, and easily collaborate on historical documents. In your browser* [online]. Innsburck : READ-COOP [cit. 2022-12-10]. Dostupné na: https://lite.transkribus.eu/

Transkribus skopírovať na ostatné strany rukopisu. V prípade Csákósovho katalógu však táto automatizácia nebola možná, keďže ide o rukou narysované tabuľky, ktoré sa na jednotlivých stranách líšia svojou pozíciou aj vnútornou štruktúrou buniek, ktoré ich autor prispôsobil vpisovanému obsahu. Segmentáciu už potom bolo možné dokončiť automaticky – *Transkribus* v preddefinovaných rámcoch tabuľky rozoznal a označil hranice riadkov textu. Manuálne bolo potrebné ich iba korigovať, pričom dôležité bolo len zachytiť začiatok a koniec tzv. základnej línie textu (baseline), resp. zmenu jej orientácie. Časová náročnosť úplnej segmentácie jednej dvojstrany bola priemerne 20 minút (t. j. 10 minút na jednu stranu dokumentu).



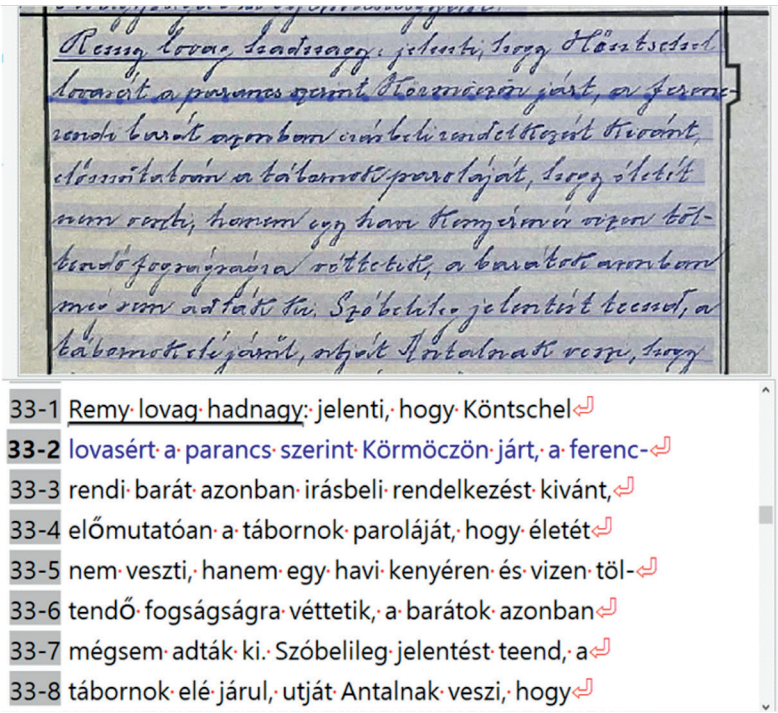
Obrázok 23 Vyznačené textové rámce a hranice riadkov s určením poradia čítania na dvojstránke Csákósovho rukopisného katalógu. Zdroj: *Transkribus*.

PRÍPRAVA MODELU NA AUTOMATICKÚ TRANSKRIPCIU

Na prípravu modelu na automatickú transkripciu bolo potrebné pripraviť bezchybný prepis vzorky textu (označovaný ako Ground Truth). Pre takúto vzorku sa odporúča pripraviť prepis rukopisu v rozsahu približne 15 000 slov.[16] V prípade prvého modelu pre Csákósov katalóg som použil 29 snímok obsahujúcich strany 1 – 53 z prvého korpusu

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

katalógu. Kľúčovým je presný prepis originálu spočívajúci v priradení konkrétnych alfalnumerických znakov, ktoré exaktne zodpovedajú originálnemu rukopisu, t. j. kopírujú aj všetky jeho omyly a nepresnosti, pričom akékoľvek odchýlenie od originálu sa môže neskôr prejaviť na chybovosti modelu. Ďalším dôležitým aspektom je pomerné zastúpenie jednotlivých znakov, resp. variácií jednotlivých znakov. Akékoľvek odchýlky v rukopise (napr. zmena štýlu písania, častý výskyt autorských korektúr, hromadné uvádzanie číselných údajov) môže takisto negatívne ovplyvniť výslednú úspešnosť modelu. Preto je dôležité pri príprave Ground Truth vzorky zvoliť čo najviac typické, reprezentatívne strany z 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

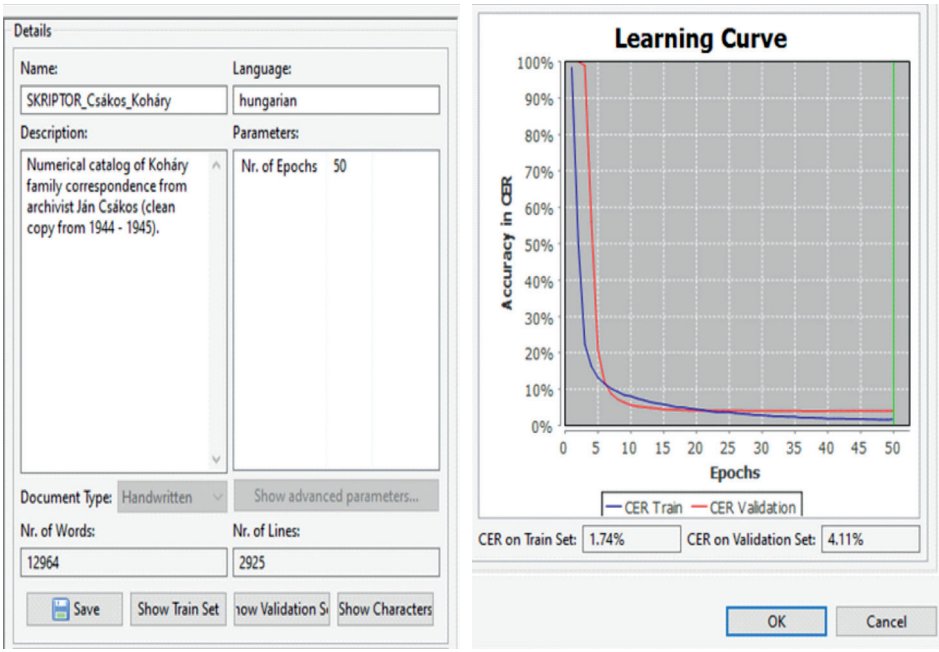
Obrázok 24 Ukážka prepisu rukopisného originálu v *Transkribe*. Zdroj: *Transkribus*.

Vzorku Ground Truth som rozdelil v odporúčanom 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). Experimentom som overil, že zvyšovanie týchto cyklov nemá takmer žiadny vplyv na výsledok, t. j. na výslednú úspešnosť modelu pre automatickú transkripciu (porov. aj experiment P. Maliniaka v 2. kapitole tejto knihy pri trénovaní modelu pre rukopis postily Izáka Abrahamidesa). 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 (podrobnejšie pozri stať D. Katuščáka *Tvorba modelu transkripcie* v úvodnej kapitole tejto knihy). Na konci tohto procesu máme k dispozícii model na automatický prepis rukopisu J. Csákósa aj s jeho základnými charakteristikami (obrázok 25).

VYHODNOTENIE ÚSPEŠNOSTI MODELU A JEHO ĎALŠIE ZDOKONAĽOVANIE

V charakteristike modelu č. 1 sú 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).



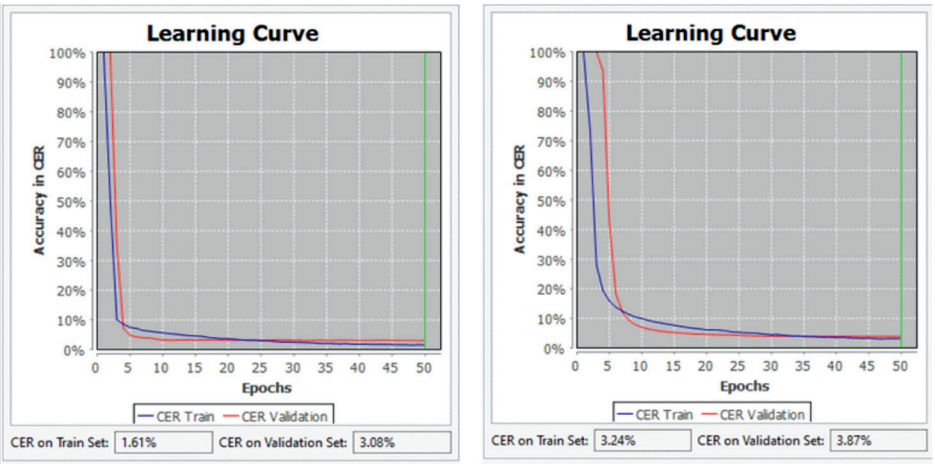
Obrázok 25 Charakteristika modelu č. 1 na automatickú transkripciu Csákósovho katalógu. Zdroj: *Transkribus*.

Na obrázku 25 je v grafickej podobe znázornený 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 *Transkribu* zvládnuť „prečítanie“, resp. automatický prepis ľubovoľnej strany skúmaného rukopisu. Konkrétne údaj hovorí o tom, že 95,89 % 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. Lepšie hodnoty okolo 1 – 2 % CER sa zvyknú dosahovať len pri modeloch pre historické tlačoviny.[17]

Vzhľadom na to, že originálny rukopis Csákósovho katalógu predstavuje počtom strán mimoriadne rozsiahly celok, rozhodol som sa pokračovať v zdokonaľovaní modelu na jeho automatickú transkripciu rozširovaním vzorky Ground Truth. Prácu manuálneho prepisu nových strán rukopisu už výrazne uľahčil existujúci model č. 1, s pomocou ktorého som po segmentácii textu spustil automatickú transkripciu ďalších 28 digitalizátov Csákósovho katalógu, ktoré obsahovali snímky strán 54 – 105. Digitálny text, ktorý som dostal, bol relatívne čitateľný a zrozumiteľný, obsahoval však chyby najmä v interpunkcii (chýbajúca alebo naopak nadbytočná bodka, čiarka, dvojbodka a pod.), v diakritike (krátka samohláska namiesto dlhej, resp. naopak), v čísliciach, ale aj pri niektorých písmenách (badateľná bola najmä tendencia automatického prepisu „ll“ namiesto správneho „k“ a pod.). Všetky chyby som na základe digitalizátu originálneho textu korigoval a zaradil do novej vzorky Ground Truth, z ktorej išlo do cvičného súboru 14 194 slov.

Na zdokonaľovanie modelu sa ponúkali dve možnosti: spojiť obe vzorky Ground Truth do jednej, z ktorej sa vyčlení pre cvičný súbor spolu 24 683 slov a vytrénovať nový model, alebo použiť model č. 1 ako základný model (Base Model) pre nový model vytrénovaný iba z novej vzorky Ground Truth. Experiment nám potvrdil, že nový model (č. 2) vytrénovaný s použitím základného už existujúceho modelu pre Csákósov katalóg má lepšie charakteristiky (nižšiu chybovosť CER a WER) ako nový model (č. 3) vytrénovaný na celej vzorke 24 683 slov bez použitia základného modelu (pozri obrázok 26).

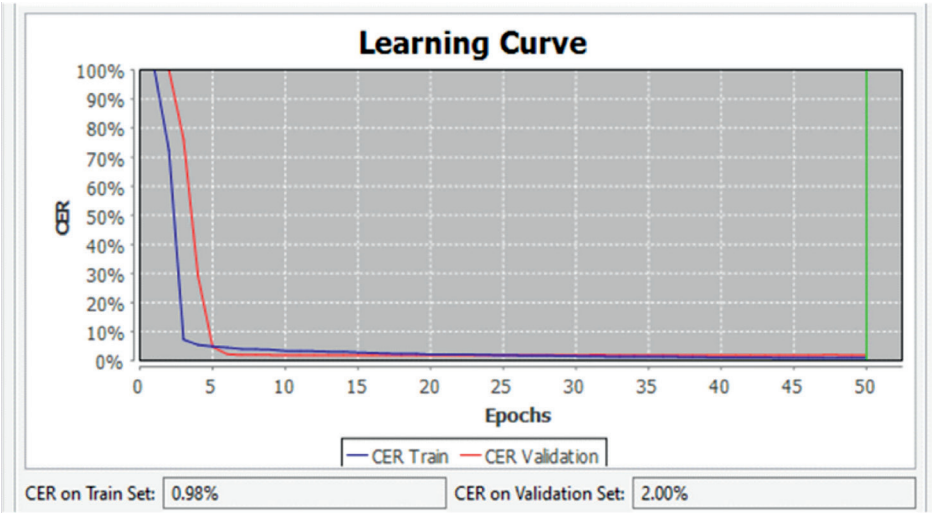


Obrázok 26 Porovnanie rozšírených modelov na automatickú transkripciu Csákósovho katalógu. Vľavo model č. 2, vpravo model č. 3. Zdroj: *Transkribus*.

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

V prvom prípade ukazujú charakteristiky nového modelu č. 2 chybovosť CER 1,61 % na cvičnom súbore a 3,08 % na overovacom súbore. Ak sa tieto údaje porovnajú s chybovosťou modelu č. 1, možno konštatovať, 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 č. 3 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 model.

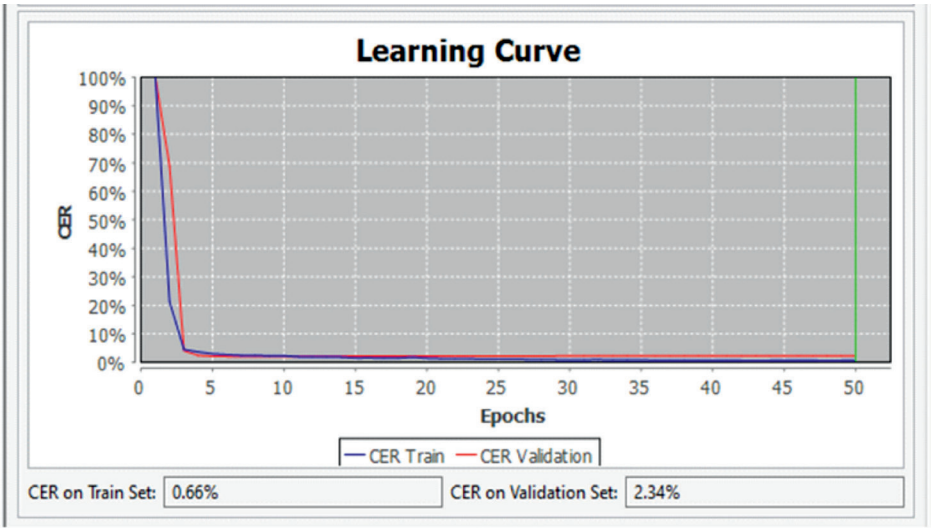
Otázka je, či možno dosiahnuť ďalšie zdokonaľovanie modelu pridávaním nových vzoriek Ground Truth, resp. dokedy má takéto zdokonaľovanie modelu zmysel. Na jej zodpovedanie bolo potrebné vykonať ďalšie experimenty. Použil som pritom postup opísaný pri trénovaní modelu č. 2. To jest model č. 2 som aplikoval na automatickú transkripciu ďalšej vzorky digitalizátov v rozsahu 52 strán (strany 106 – 157) rukopisu. Po ich nevyhnutnej korekcii na úroveň Ground Truth som spustil trénovanie modelu č. 4 s využitím základného modelu č. 2. Do cvičného súboru pre nový model bolo zaradených 16 307 slov.



Obrázok 27 Charakteristika modelu č. 4 na automatickú transkripciu Csákósovho katalógu. Zdroj: *Transkribus.*

Ako vidno z charakteristík tohto nového modelu č. 4 (obrázok 27), úroveň chybovosti poklesla oproti modelu č. 2 o ďalších 1,08 percentuálneho bodu na 2,00 %. Na prvý pohľad to vedie k optimistickému záveru, že rozširovaním vzorky Ground Truth a pridávaním slov do cvičného súboru možno znižovať chybovosť na limitné úrovne pod 2 %, resp 1 %. Uvedené by platilo, ak by originálny rukopis mal konzistentnú podobu s pomerným zastúpením znakov, resp. slov na jednotlivých stranách. V realite sa však s takýmito ideálnymi podmienkami nestretávame. Overil som si to aj v prípade Csákósovho katalógu, keď som sa rozhodol pokračovať v zdokonaľovaní modelu na jeho automatickú transkripciu pridaním ďalšej vzorky Ground Truth z nadväzujúcich strán 158 – 191 rukopisu, t. j. v rozsahu 34 strán. Vďaka tomu som mohol na základe modelu č. 4.

(Base Model) s použitím nového cvičného súboru obsahujúceho 8 466 slov vytrénovať ďalší model č. 5. Jeho charakteristiky však už v prípade parametrov chybovosti neboli také jednoznačné (pozri obrázok 28).

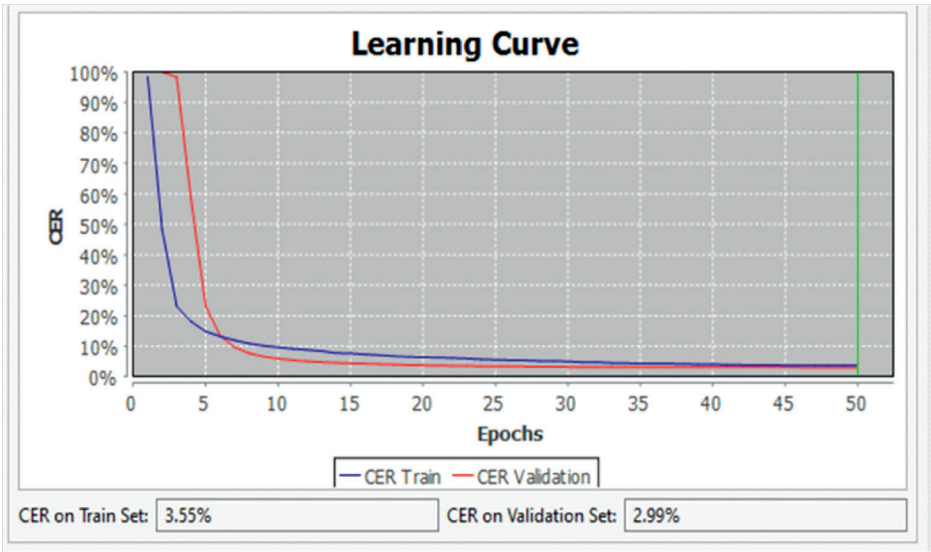


Obrázok 28 Charakteristika modelu č. 5 na automatickú transkripciu Csákósovho katalógu. Zdroj: *Transkribus.*

V prípade miery chybovosti na cvičnom súbore možno konštatovať ďalší pokles. Tento ukazovateľ teda kontinuálne potvrdzuje zdokonaľovanie modelu, keďže u každého ďalšieho (rozšíreného) modelu vykazuje nižšiu úroveň: model č. 1 – 1,74 %, model č. 2 – 1,61 %, model č. 4 – 0,98 % a napokon model č. 5 – 0,66 %. Toto však už neplatí pri sledovanom parametri chybovosti na overovacom súbore, ktorý by mal byť určujúcim na vyhodnotenie úspešnosti modelu. V prípade modelu č. 5 tento parameter vykázal nárast oproti predchádzajúcemu modelu č. 3 o 0,34 percentuálneho bodu. Tento výsledok však nemožno považovať za zlyhanie modelu, ani neoprávňuje konštatovať, že som narazil na limity samotnej platformy automatickej transkripcie. V skutočnosti to odzrkadľuje rôznorodosť rukopisu, kvality jeho digitalizátov a v neposlednom rade výber strán pre cvičný a v tomto konkrétnom prípade najmä pre overovací súbor. Na týchto stranách sa totiž vo zvýšenej miere nachádzali číselné údaje (text obsahoval číselné výkazy a účty). Frekvencia výskytu číslic v tomto rukopise je však oproti iným znakom (písmenám) rádovo mnohonásobne nižšia. Navyše číslice vykazujú nepomerne vyššiu variabilitu spôsobu zápisu ako v prípade písmen. Platforma teda nemá dostatok vzorov na ich spoľahlivé rozlíšenie. V konečnom dôsledku to vedie k zvýšeniu ukazovateľa chybovosti na overovacom súbore. Chybovosť ovplyvňuje aj kvalita digitalizátu – vplyv na ňu majú nielen technické špecifikácie snímky, ale aj charakteristiky samotného originálu rukopisu (napr. farebná sýtosť písma).

Na porovnanie s modelom č. 5, ktorý možno už považovať za definitívny a úspešný model na automatickú transkripciu Csákósovho katalógu, som vytvoril ešte raz aj model zo všetkých vzoriek Ground Truth spoločne bez využitia základného modelu. Cvičný

súbor pre jeho potreby tak obsahoval mimoriadne rozsiahlu množinu slov v počte 48 623. Pri hodnotení tohto modelu č. 6 sa potvrdila hypotéza, ktorú som vyslovil pri modeli č. 3. To jest nižšiu mieru chybovosti možno dosiahnuť rozšírením vzorky Ground Truth s použitím pôvodnej verzie modelu ako základného – Base Modelu. Aj model č. 6 vytrénovaný bez jeho využitia dosiahol vyššiu mieru chybovosti ako model č. 5, ba dokonca aj ako model č. 4 (pozri obrázok 29).



Obrázok 29 Charakteristika modelu č. 6 na automatickú transkripciu Csákósovho katalógu. Zdroj: *Transkribus.*

MOŽNOSTI VYUŽITIA DIGITALIZOVANÉHO KATALÓGU J. CSÁKÓSA V ARCHÍVNEJ PRAXI

S využitím modelu č. 6 pre rukopis J. Csákósa možno pristúpiť k úplnej digitalizácii katalógu (rozumej prevod na digitálny text) korešpondencie Kohárovcov. V rámci príprav definitívneho modelu č. 6 som pre vzorku Ground Truth prepísal, resp. po automatickej transkripcii korigoval 191 strán z celkového počtu 4 140 strán celého katalógu. Pri príprave digitalizátov zvyšných 3 949 strán rukopisného katalógu je potrebné počítať s časovou náročnosťou manuálnej segmentácie jednotlivých strán. Pri predpokladanom intervale približne 10 minút na 1 stranu to činí približne 658 hodín čistej práce. Ďalším nákladom je cena za automatickú transkripciu textu na platforme *Transkribus*. Za automatickú transkripciu jednej snímky digitalizátu sa účtuje 1 kredit. Vzhľadom na to, že jedna snímka digitalizátu Csákósovho katalógu obsahuje dvojstranu, na transkripciu všetkých zvyšných strán je potrebných 1 975 kreditov. Pri súčasnej cenovej politike READ-COOP stojí balík 2000 kreditov 528,- €. Získanie plnohodnotného digitálneho textu jednej strany Csákósovho katalógu teda vychádza priemerne na necelých 0,14 €. V časovom rozmedzí 3 až 4 mesiacov teda môže získať archív pri vyťažení jedného pracovníka bez špeciálnych nárokov na jeho odborné archívne vzdelanie či na znalosť jazykov prameňov plne digitalizovaný text dobovej archívnej pomôcky. Aj takto

– finančne a časovo – možno vyjadriť ekonomickú úsporu, ktorú ponúka platforma *Transkribus* pri úplnej digitalizácii dobovej archívnej pomôcky v porovnaní s tým, ak by prepisoval manuálne stranu po strane odborný archívny pracovník so znalosťou jazykov historických prameňov (maďarčina, nemčina, latinčina) a skúsenosťami s čítaním historických rukopisov, prípadne aj paleografie. (Kým cena za takto manuálne prepísanú stranu rukopisu by sa dala odvodiť od výšky hodinovej mzdy dotyčného odborného archívneho pracovníka, časovú náročnosť takejto práce si netrúfam odhadnúť.)

Digitálny text možno z *Transkribu* 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). *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 digitalizácii textu vďaka automatickej transkripcii môže mať výsledok k dispozícii takmer okamžite. Tabuľkový výstupný formát (xlsx) zas umožňuje štruktúrovanie dát, resp. ich zoraďovanie napr. podľa roku, miesta, ale aj akejkoľvek ďalšej doplňujúcej informácie (napr. pôvodcu, adresáta atď.). Vďaka takémuto sprístupneniu dobovej archívnej pomôcky môže odborný archívny pracovník jednoduchšie vyhľadávať konkrétne listy a v nich konkrétne témy, resp. osoby. Takisto mu to výrazne uľahčí prácu pri vypracovaní moderných sprievodcov po fonde, resp. ďalších archívnych pomôcok.

ZÁVER

Aplikácia metódy automatickej transkripcie historických rukopisných textov prostredníctvom platformy *Transkribus* ponúka možnosť rýchlej, jednoduchej, a výhodnej digitalizácie dobových archívnych pomôcok. Vynaložené náklady vyvažuje sprístupnenie inak nespracovaného archívneho fondu pre ďalší archívny alebo historický výskum. Ako argument pre takýto záver možno použiť práve model automatickej transkripcie pre Csákósov rukopisný katalóg korešpondencie Kohárovcov. Na príklade tohto rukopisného textu možno definovať aj ideálny materiál na aplikáciu automatickej transkripcie – rozsiahly rukopisný text obsahujúci niekoľko stoviek, najlepšie tisícok strán, písaný jednou rukou v unifikovanej forme. (Na margo dodávam, že rovnako možno touto metódou previesť do digitálnej podoby aj archívne pomôcky mladšieho dáta, ktoré existujú napr. v podobe strojopisu, pri ktorých *Transkribus* úspešne konkuruje či nahrádza – najmä v prípade horšie čitateľných textov – iné nástroje OCR.) Digitalizovaný text možno po obsahovej analýze, nevyhnutných korektúrach a formálnych úpravách publikovať v digitálnej i tlačenej verzii, resp. na základe neho vypracovať ďalšie pomôcky – registre (menný, miestny, vecný, chronologický, pôvodcov, adresátov atď.) Ďalším krokom bude selekcia listov podľa pôvodcu. Ako už bolo uvedené pri charakteristike fondu, ide o mimoriadne rozsiahly súbor 6 632 listov, čo oprávňuje predpokladať, že sa takto otvorí aj cesta na vytvorenie modelu automatickej transkripcie rukopisu jednotlivých

členov rodiny Koháryovcov, resp. iných dôležitých historických postáv, ktorých listy sa nachádzajú v tomto fonde vo významnom počte (odhadom by to malo byť minimálne 50 listov). Samotný model pre Csákósov katalóg korešpondencie Koháryovcov ponúka pre seba aj ďalšie využitie. Pripomínam, že J. Csákós bol mimoriadne bádateľsky činný mestský archivár v Bratislave. Model vyvinutý na prepis jeho katalógu korešpondencie môže byť aplikovateľný na akýkoľvek rukopisný text, ktorý napísal. Overenie uvedených hypotéz bude ďalším krokom pri testovaní funkčnosti a využiteľnosti platformy *Transkribus* v podmienkach našich pamäťových inštitúcií (archívov).

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

Download Transkribus. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-12-20]. Dostupné na: https://readcoop.eu/transkribus/download/

FONÓD, Zoltán (ed.): *A cseh/szlovákiai magyar irodalom lexikona 1918 – 2004.* 2. oprav. vyd., Bratislava : Madách-Posonium, 2004. 480 s.

FRANCOVÁ, Zuzana: Jozef Ján Csákós (1883 – 1957), bývalý kurátor múzea. In: *Zborník Múzea mesta Bratislavy*, roč. 29. Bratislava : Múzeum mesta Bratislavy, 2017, s. 239 – 244.

GOMBOS, János: A Koháry-Coburg család besztercebányai családi levéltára. In: *Levéltári Szemle*, roč. 43, č. 2, 1997, s. 35 – 41.

CHALUPECKÝ, Ivan: K problematike štátnych a cirkevných archívov na Slovensku. In: *Z minulosti Spiša. Ročenka Spišského dejepisného spolku v Levoči – XXIV,* roč. 2016. Levoča : Spišský dejepisný spolok Levoča, 2016, s. 185 – 198.

KOMJÁTI, Zoltán Igor: „… Az mit hallottam, kötelességem szerint akarám Nagyságodnak tudtára adnom…" Adalékok a híráramlás és a hírhálózat történetéhez Felső-Magyarországon Koháry István füleki főkapitány levelezésének tükrében (1672 – 1682). In: *Fons (Forráskutatás és Történeti Segédtudományok)*, roč. XVII, č. 1, 2010, s. 113 – 140.

KOPČAN, Vojtech: *Turecké nebezpečenstvo a Slovensko*. Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 1986. 222 s.

LENGYELOVÁ, Tünde: „Ja potrebujem peniaze viac, ako tvoja milosť": Korešpondencia Judity Balassovej so synom Štefanom II. Kohárym. In: *„Za Boha, kráľa a vlasť!": Kohároyovci v uhorských dejinách. Zborník príspevkov z odborného seminára, ktorý sa konal 23. – 24. septembra 2015 v Múzeu vo Svätom Antone*. ed. M. Ďurianová, Zvolen : Národné lesnícke centrum, 2016, s. 9 – 19.

MATUNÁK, Michal: *Krupinskí hradní kapitáni*. ed. Miroslav Lukáč, Krupina : Kultúrne centrum a Múzeum Andreja Sládkoviča v Krupine v spolupráci s Mestom Krupina, 2011, [xli], 95 s.

MUEHLBERGER, Guenter et al.: Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: *Journal of Documentation*, vol. 75, no. 5, 2019, pp. 954 – 976.

OTRUBA, Štefan: *Štátny archív v Banskej Bystrici: sprievodca po archívnych fondoch II.* Bratislava : Slovenská archívna správa, 1969. 284 s.

RIGELE, Alojz: Portrét Jozefa Jána Csákósa. Litografia. 1923. Galéria mesta Bratislavy [cit. 2022-12-12]. Dostupné na: https://www.webumenia.sk/dielo/SVK:GMB.C_13385

STRÖBEL, Phillip – CLEMATIDE, Simon: *Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution*

Images [online]. Utrecht : Digital Humanities 2019. Posted at the Zurich Open Repository and Archive, University of Zurich [cit. 2021-08-26]. Dostupné na: https://doi.org/10.5167/uzh-177164

Štátny archív Banská Bystrica, fond Koháry – Coburg (1241) 1321 – 1945.

The ScanTent: Professional scanning with your smartphone. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-12-10]. Dostupné na: https://readcoop.eu/scantent/

Transkribus Lite: Automatically transcribe, comfortably edit, and easily collaborate on historical documents. In your browser [online]. Innsburck : READ-COOP [cit. 2022-12-10]. Dostupné na: https://lite.transkribus.eu/

Zákon č. 395/2002 Z. z. o archívoch a registratúrach a o doplnení niektorých zákonov v znení neskorších predpisov. Dostupné tiež na: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2002/395/

KAPITOLA 4

AUTOMATICKÁ TRANSKRIPCIA PROTOKOLOV Z KANONICKÝCH VIZITÁCIÍ FARNOSTÍ ZVOLENSKÉHO ARCIDIAKONÁTU Z POLOVICE 18. STOROČIA V PLATFORME TRANSKRIBUS

Patrik Kunec

Univerzita Mateja Bela v Banskej Bystrici; Filozofická fakulta; Katedra histórie
E-mail: patrik.kunec@umb.sk

ABSTRAKT

Obsahom štúdie je predstaviť metódy a postupy pri príprave automatickej transkripcie historického prameňa v platforme *Transkribus* v rámci riešenia úloh aplikovaného výskumu *SKRIPTOR*. Spracovávaným prameňom je súbor po latinsky písaných protokolov z kanonických vizitácií farností Zvolenského arcidiakonátu z rokov 1754 – 1755/1756. V úvodnej časti je opísaný historický kontext vytvorenia prameňa a jeho význam, priblížená je aj jeho vnútorná štruktúra a stručný obsah. Následne je vysvetlený metodický postup spracovania prameňa v platforme *Transkribus* pre potreby automatickej transkripcie jeho obsahu. Určitý hendikep vybraného prameňa na účely automatickej transkripcie predstavuje fakt, že ide o súbor textových dokumentov vytvorených rozličnými autormi. Práve variabilita rukopisov, aj keď sú si do veľkej miery podobné, do istej miery komplikuje vytvorenie takého softvérového nástroja (modelu), ktorý by dokázal tieto texty automaticky transkribovať s čo najmenšou chybovosťou. Hoci boli modely vytvárané na vzorke s malým rozsahom prepísaných strán, dosiahnutá miera ich chybovosti v rozpoznávaní znakov v overovacom súbore dosahovala hodnoty okolo 7 %.

Kľúčové slová: protokoly kanonických vizitácií; historický prameň; Zvolenský arcidiakonát; 18. storočie; automatická transkripcia; *Transkribus*

ABSTRACT

Automatic transcription of the protocols from canonical visitations of the parishes of the Archdeaconry of Zvolen from the mid-18th century in the Transkribus platform

The content of the study is the presentation of methods and procedures for the preparation of automatic transcription of a historical source in the *Transkribus* platform as part of solving the tasks of applied research SKRIPTOR. The processed source is a set of the protocols, written in Latin, from canonical visitations of the parishes of the Archdeaconry of Zvolen from the years 1754 – 1755/1756. In the

introductory part, the historical context of creation of the source and its meaning are described, its internal structure and brief content are also described. Subsequently, the methodical procedure of processing the source in the *Transkribus* platform for the needs of automatic transcription of its content is explained. A certain handicap of the selected source for the purposes of automatic transcription represents the fact that it is a set of text documents created by different authors. It is the variability of the manuscripts, even if they are largely similar, that complicates, to a certain extent, the creation of such a software tool (model) that could automatically transcribe these texts with as little error rate as possible. Although the models were created on the scale of a small sample of transcribed pages, their error rate in character recognition in the verification set reached a value of around 7 %.

Keywords: protocols of canonical visitations; historical source; Archdeaconry of Zvolen; 18th century; automatic transcription;*Transkribus*

ÚVOD

Na skúmanie náboženských pomerov a cirkevnej štruktúry v období raného novoveku máme na našom území zachovaných a sprístupnených niekoľko druhov dobových písomných prameňov, ale najlepšiu výpovednú hodnotu pre historikov majú tzv. zápisnice z kanonických vizitácií (tiež protokoly z kanonických vizitácií, alebo vizitačné protokoly), ktoré sa objavujú práve od polovice 16. storočia. Slovné spojenie kanonická vizitácia (lat. *visitatio canonica*) má v historiografii dva významy. V prvom, historicky presnejšom či adekvátnejšom význame ide o pravidelne konané kontrolné návštevy jednotlivých farností alebo rôznych cirkevných inštitúcií v hraniciach diecézy alebo arcidiecézy, ktorých konanie nariaďoval biskup, arcibiskup, prípadne samotný panovník. V druhom význame, menej historicky a obsahovo presnom, toto spojenie označuje písomné záznamy vizitátora (alebo viacerých vizitátorov) o zisteniach alebo výsledkoch jeho kontrolnej návštevy. Terminologicky vhodnejšie pomenovanie pre kanonické vizitácie ako písomný prameň úradnej povahy by malo byť zápisnice z kanonických vizitácií, prípadne vizitačné protokoly (tieto dva termíny používam v prítomnom texte).

Hoci patria vizitačné protokoly z obdobia novoveku svojím obsahom k významným prameňom poznania nielen cirkevných, ale aj sociálnych a do istej miery aj hospodárskych a kultúrnych dejín, za posledné polstoročie sa ich skúmaniu a sprístupňovaniu v slovenskej historiografii venovala skôr malá pozornosť.[1] Kompletných edícií parciálnych zápisníc z kanonických vizitácií z rôznych období a lokalít je v našom prostredí publikovaných len niekoľko,[2] veľmi skromne je analyzovaný a zhodnotený ich obsah

1 LOPATKOVÁ, Zuzana: Edície zápisníc z kanonických vizitácií z obdobia novoveku. In: *Metodologické limity historického prameňa.* eds. M. Kohútová – Z. Lopatková, Kraków ; Trnava : Towarzystwo Słowaków w Polsce ; Filozofická fakulta Trnavskej univerzity v Trnave, 2014, s. 38.

2 ŠIMONČIČ, Jozef – KARABOVÁ Katarína (eds.): *Kanonické vizitácie Dunajeckého dekanátu v Spišskom biskupstve z roku 1832*. Kraków : Towarzystwo Słowaków w Polsce, 2015. 848 s.; KLEMENT, Martin: *Pramene k dejinám Rímskokatolíckej cirkvi, farnosti Bošáca (Kánonické vizitácie z rokov 1797, 1816 a 1829)*. Trnava : Ivona Matúšová, 2017. 248 s.; LOPATKOVÁ, Zuzana (ed.): *Visitationes Canonicae Archidiaconatus Posoniensis 1694/1695.* Trnava : Filozofická fakulta Trnavskej univerzity v Trnave, 2020. CD-ROM.

a význam aj po metodologickej stránke.[3] Vzhľadom na ich dôležitosť pre poznanie sociálnych a kultúrno-náboženských pomerov v malých komunitách je to dosť prekvapujúci stav. Čoraz častejšie sa zápisnice z kanonických vizitácií využívajú pri písaní monografií o dejinách obcí alebo miest, prípadne sa prostredníctvom ich informačného obsahu analyzujú problémy náboženskej a sociálnej koexistencie rôznych vierovyznaní. Príkladom využitia protokolov z kanonických vizitácií v tejto oblasti výskumu môže byť originálna práca historikov Petra Zubku a Petra Žeňucha, v ktorej na podklade vizitačných protokolov sledujú minoritné náboženské spoločenstvá evanjelikov a gréckokatolíkov na východnom Slovensku.[4] Potvrdenie konštatovania, že protokoly z kanonických vizitácií poskytujú neoceniteľné informácie aj pre dejiny kultúry, dokladá aj nedávno publikovaná trojzväzková monografia Dariny Múdrej o hudobnej produkcii v cirkevnom prostredí na základe informácií z vizitačných protokolov.[5] Je zrejmé, že postupne sa zápisnice z kanonických vizitácií stanú čoraz atraktívnejším prameňom pre početných bádateľov, ktorí by určite uvítali, keby boli tieto historické dokumenty sprístupnené moderným spôsobom – teda v digitalizovanej podobe na internete.

CHARAKTERISTIKA ZÁPISNÍC Z KANONICKÝCH VIZITÁCIÍ AKO HISTORICKÉHO PRAMEŇA

Snaha cirkevných inštitúcií spoznať reálny stav náboženských pomerov v krajine viedla ku kontrole kňazov a farností už vo vrcholnom a neskorom stredoveku. Prvé správy o konaní kanonických vizitácií z uhorského územia sa viažu k roku 1397.[6] Záväzné nariadenie o konaní pravidelných vizitácií biskupmi alebo arcibiskupmi priniesli až uznesenia Tridentského koncilu (konkrétne išlo o závery z jeho 24. zasadania zo dňa 11. 11. 1563). Vizitácie farností boli jedným z rekatolizačných opatrení, ktoré mali prispieť k obnove rímskokatolíckej cirkvi. Prvé kanonické vizitácie vykonal v Uhorsku ešte v čase konania Tridentského koncilu ostrihomský arcibiskup Mikuláš Oláh (vo funkcii v rokoch 1553 – 1568). Aby lepšie spoznal pomery v územne rozsiahlom arcibiskupstve, rozposlal v rokoch 1559 – 1564 do jednotlivých arcidiakonátov vizitátorov, pričom niekoľko záznamov (konkrétne z desiatich arcidiakonátov Ostrihomskej arcidiecézy) sa zachovalo do súčasnosti.[7] Tieto prvé kanonické vizitácie sa zameriavali na osobu kňaza, na počet veriacich a ich rozdelenie z hľadiska príslušnosti ku katolíckej cirkvi alebo k protestantským denomináciám a tiež na aktuálny stav budovy kostola. Ich rozsah bol skôr stručný. Jednotná podoba kanonických vizitácií, ako aj zápisníc z kontrolných návštev bola v Uhorsku prediskutovaná a definovaná až na všeobecných synodách, ktoré sa konali v rokoch 1611 a 1629 v Trnave. Pokyn na uskutočnenie vizitácií z roku

3 KYSEĽOVÁ, M.: Kanonické vizitácie na Spiši. In: *Slovenská archivistika*, roč. XVIII, 1983, č. 2, s. 110 – 128 a LOPATKOVÁ, Zuzana: *Kanonické vizitácie 16. – 18. storočia v slovenských dejinách.* Trnava : Filozofická fakulta Trnavskej univerzity v Trnave, 2021, 64 s.

4 ZUBKO, Peter – ŽEŇUCH, Peter: *Barkóciho vizitácia Šarišského archidiakonátu (1749): rímskokatolíci, gréckokatolíci a evanjelici podľa latinskej vizitácie*. Bratislava : Slavistický ústav Jána Stanislava pri SAV, 2017, 190 s.

5 MÚDRA, Darina: *Topografia hudby klasicizmu na Slovensku z pohľadu kanonických vizitácií.* Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 2019. 1318 s.

6 ŠIMONČIČ, J. – KARABOVÁ, K.: *Kanonické vizitácie Dunajeckého dekanátu*..., s. 7.

7 LOPATKOVÁ, Z.: *Kanonické vizitácie 16. – 18. storočia v slovenských dejinách*, s. 8.

1611 nebol realizovaný v praxi a vizitácie nariadené arcibiskupom Petrom Pázmáňom v roku 1629 sa realizovali iba pomaly, pričom ukončené boli až v roku 1634.[8] Ďalšie vizitácie sa v Ostrihomskom arcibiskupstve konali v rokoch 1674 a 1694. V 18. storočí sa konali vizitácie jednotlivých arcidiakonátov arcibiskupstva v rokoch 1754 – 1756, v roku 1761 a napokon v rokoch 1781 – 1783.

Vykonanie vizitácie v rámci celého biskupstva alebo jeho časti (zvyčajne arcidiakonátov) nariaďoval arcibiskup alebo biskup, výnimočne tak mohol urobiť aj panovník. Podľa statusu či postavenia vizitátora sa vizitácie rozdeľujú na biskupské, arcidiakonálne a dekanské.[9] V 18. a 19. storočí biskupi zvyčajne poverovali konaním vizitácií arcidiakonov alebo dekanov, prípadne príslušníkov väčších kapitúl. Vykonanie vizitácie bolo nariadené písomne a často sa znenia týchto textov v úvode protokolov z kanonických vizitácií zachovali. Vo výzve realizovať vizitácie sa zdôrazňovala úloha vizitátora zistiť a opísať reálny stav náboženských komunít vo vymedzenom cirkevno-správnom okrsku. Cieľom bolo dosiahnuť nápravu nedostatkov, ktoré boli počas vizitácie zistené. Priebeh a časový plán vizitácie mohla i nemusela upravovať biskupská inštrukcia. O plánovanej návšteve vizitátora vo farnostiach boli ich predstavitelia vopred informovaní, od polovice 18. storočia už v predstihu dostali aj zoznam prípadných otázok, na ktoré mali podať odpovede. V jednej farnosti sa zdržiaval vizitátor len jeden deň a z návštevy farnosti sa vytvárala zápisnica, ktorá bola neskôr prepísaná načisto a potvrdená podpisom a pečaťou vizitátora, niekedy aj zemepána či zemepánov, ktorí boli patrónmi farnosti. Potvrdené čistopisy protokolov z kanonických vizitácií boli uložené v biskupskom alebo arcibiskupskom archíve.

Vizitátor počas návštevy farnosti zaznamenal informácie o osobe farára (prípadne aj kaplána a kostolníka), stručne opísal náboženskú príslušnosť farníkov (vo vizitačných protokoloch sú záznamy aj o protestantských veriacich), priblížený bol aj vzťah zemepána (zemepánov) s patronátnym právom ku konkrétnej farnosti. Zisťovala sa aj hospodárska stránka fungovania farnosti, teda majetok a príjmy kňaza, výška poplatkov za jednotlivé cirkevné obrady a úkony, prípadne aj výška a podoba naturálnych dávok farníkov. Ak mal kňaz podlžnosti či finančné záväzky, tieto boli taktiež do zápisnice zaznamenané. V časovo mladších vizitačných protokoloch boli zaznamenané aj majetkové pomery farníkov, najmä rozsah im zverených poľností. Súčasťou zápisnice z kanonickej vizitácie bol aj opis chrámu a jeho vnútorného vybavenia, prípadne aj opis budovy fary. V 18. storočí už vizitátori venovali pozornosť aj škole a učiteľovi, ktorý v nej pôsobil, farárovej knižnici, prípadne pôsobeniu pôrodných babíc.

Protokoly z kanonických vizitácií sú z dôvodu bohatstva informácií a ich vysokej objektívnosti dôležitým historickým prameňom. Historikom poskytujú cenné fakty o náboženskom, spoločenskom, hospodárskom a kultúrnom charaktere dobovej spoločnosti. Vizitačné protokoly sú bohatým zdrojom dát aj pre výskum dejín umenia (najmä architektúry, maliarstva a umeleckého remesla), dejín hudby, ale aj dejín školstva, knižnej kultúry či zdravotníctva. Z informácií v nich obsiahnutých čerpajú bádatelia aj v oblasti historickej demografie, štatistiky, topografie, lingvistiky a etnológie.

8 LOPATKOVÁ, Z.: Edície zápisníc z kanonických vizitácií..., s. 42.

9 BOŠANSKÝ, Martin: *Kánonické vizitácie Nitrianskej diecézy do roku 1831.* Nitra : Kňazský seminár sv. Gorazda, 2018, s. 20.

CHARAKTERISTIKA PROTOKOLOV Z KANONICKEJ VIZITÁCIE ZVOLENSKÉHO ARCIDIAKONÁTU Z ROKOV 1754 – 1756

Pri riešení úloh v rámci projektu SKRIPTOR som sa aj na základe diskusií s ďalšími riešiteľmi projektu z Katedry histórie Filozofickej fakulty Univerzity Mateja Bela v Banskej Bystrici rozhodol pre výber po latinsky písaných prameňov úradnej povahy z 18. storočia. Pri zvažovaní vhodného typu prameňa (zápisnice zo stoličných kongregácií, mestské knihy, pramene cirkevnej proveniencie) napokon prevážil názor vybrať na spracovanie v platforme *Transkribus* texty niekoľkých vizitačných protokolov z obdobia 18. storočia. Argumentom ich výberu bola do veľkej miery identická štruktúra protokolov z vizitácií jednotlivých farností, opakujúca sa slovná zásoba a tiež relatívne dobrá čitateľnosť čistopisov týchto protokolov.

V archíve Banskobystrickej diecézy so sídlom v Badíne sa nachádza obsiahly fond vizitačných protokolov z obdobia rokov 1754 – 1829 (Archív Biskupstva v Banskej Bystrici, fond Zbierka kanonických vizitácií; zaužívaný skrátený zápis: ABBB, fond ZKV).[10] Banskobystrické biskupstvo vzniklo rozhodnutím panovníčky Márie Terézie spolu s ďalšími dvomi – Rožňavským a Spišským biskupstvom, a to vyčlenením sa z územne rozsiahleho Ostrihomského arcibiskupstva. Táto zmena bola súčasťou jej cirkevných reforiem, ktoré mali najmä racionalizovať cirkevnú správu a v neposlednom rade podriadiť náboženské záležitosti kontrole štátu. O návrhu panovníčky rokovala počas vianočných sviatkov roku 1775 v Bratislave osemčlenná cirkevná komisia, ktorá jej zámer odobrila. Panovníčka vznik nových biskupstiev oznámila začiatkom roku 1776. Ich vznik postupne potvrdil bulami pápež Pius VI., v prípade Banskobystrického biskupstva tak učinil bulou *Regalium principium* zo dňa 13. marca 1776.[11] S vytváraním jednotlivých súčastí biskupskej administratívy vznikol aj biskupský archív. Je pravdepodobné, že dokumenty, ktoré vznikli pred založením biskupstva, sa do jeho archívu dostali až dodatočne z rôznych cirkevných inštitúcií. Pokiaľ ide o protokoly z kanonických vizitácií, spred roku 1776 sa do biskupského archívu v Banskej Bystrici dostali iba vizitačné protokoly z rokov 1754 – 1756, a to pravdepodobne z archívu Ostrihomského arcibiskupstva. Protokoly pochádzajú z farností nachádzajúcich sa na území Banskobystrickej diecézy v hraniciach, aké získala v čase svojho vzniku, teda v roku 1776 (nová diecéza obsiahla celú Zvolenskú a Turčiansku stolicu, ako aj severnú časť Tekovskej a dva okresy Nitrianskej stolice). Zaujímavý je najmä vizitačný protokol Zvolenského arcidiakonátu z rokov 1754 – 1756, v ktorom sú zápisnice aj z farností miest Banská Bystrica, Brezno a Zvolen.[12] Práve tento prameň som si vybral na spracovanie v projekte SKRIPTOR.

Všeobecnú vizitáciu farností Ostrihomského arcibiskupstva nariadila uhorská kráľovná Mária Terézia v máji roku 1752. Iba rok predtým sa stal arcibiskupom Mikuláš Čáki (aj Csáky, vo funkcii v rokoch 1751 – 1757). Panovníčkino nariadenie o vykonaní parciálnych vizitácií preniesol na nižšie postavených cirkevných hodnostárov, najmä na kanonikov

10 Informácia získaná z oficiálnej internetovej stránky Banskobystrickej diecézy [cit. 2022-11-25]. Dostupné na: https://bbdieceza.sk/diecezny-archiv/

11 Informácie o založení Banskobystrického biskupstva poskytuje internetová stránka diecézy [cit. 2022-11-25]. Dostupné na: https://bbdieceza.sk/banskobystricka-dieceza/historia-diecezy/zalozenie-banskobystrickej-diecezy/; Stručné informácie o vzniku diecézy ponúka aj MÚDRA, D.: *Topografia hudby klasicizmu na Slovensku*, s. 16.

12 ABBB, fond ZKV, Protokoly z kanonických vizitácií z roku 1754, signatúra CV8.

významnejších kapitúl. Vizitátorom vo farnostiach Zvolenskej stolice (a zároveň aj arcidiakonátu) bol sedmohradský kanonik gróf Jozef Baťán (aj Batthyány) z Németujváru (mestečko Güssing v rakúskom Burgenlande).[13] Od roku 1755 pôsobil aj ako prepošt Bratislavskej kapituly. Ním konaná kanonická vizitácia farností Zvolenského arcidiakonátu bola realizovaná v jesenných mesiacoch roku 1754, ale s najväčšou pravdepodobnosťou pokračovala bez jeho priamej prítomnosti aj v priebehu roku 1755. Svedčí o tom aj datovanie zápisnice z kanonickej vizitácie farnosti v Banskej Bystrici, ktoré je uvedené 1. decembrom roku 1755, a tiež fakt, že definitívnu verziu protokolov z vizitácie Zvolenského arcidiakonátu Baťán dokončil až začiatkom roku 1756. Dokladá to nielen dátum uvedený na titulnom liste, ale aj jeho list arcibiskupovi Čákimu z 2. februára 1756, v ktorom sa mu ospravedlňoval za oneskorenie vytvorenia finálnej verzie vizitačných protokolov.[14]

Parciálne vizitačné zápisnice sú zviazané v pevnej koženej väzbe, väzba je bez potlače a bez nápisu. Zviazaný dokument nemá žiadnu signatúru ani vlastnícku pečiatku inštitúcie alebo jednotlivca. Ide o zbierku vizitačných protokolov, ktoré sú určite čistopismi pôvodných verzií zápisníc z kanonických vizitácií, pretože v ich texte absentujú škrty a opravy. Niektoré časti celej zbierky sú číslované foliáciou (teda po jednotlivých listoch), v rámci celého súboru protokolov ale kontinuálna foliácia absentuje (z tohto dôvodu odkazujem ďalej na čísla jednotlivých zdigitalizovaných snímok tohto prameňa). Niektoré vizitačné protokoly – napríklad miest Banská Bystrica a Brezno (s ich filiálnymi farnosťami) sú akoby len priložené k ostatným. Protokoly väčšiny farností arcidiakonátu sú rozdelené do dvoch skupín podľa dvoch dištriktov tejto cirkevno-správnej jednotky – Horného a Dolného dištriktu. Prepisy týchto protokolov sú urobené dvomi rôznymi pisárskymi rukami.

Vizitačný protokol obsahuje aj prepisy ďalších dokumentov. Ako prvý je uvedený už spomínaný osobný list kanonika Jozefa Baťána arcibiskupovi Čákimu datovaný dňa 2. februára 1756 v Bratislave. Po ňom nasleduje viac ako štvorstranový list ostrihomského arcibiskupa Mikuláša Čákiho grófovi Jozefovi Baťánovi zo dňa 4. augusta 1754, v ktorom arcibiskup uvádza aj celé znenie nariadenia Márie Terézie o vykonaní vizitácie v jednotlivých arcidiakonátoch arcidiecézy (z mája roku 1752). Až po týchto úvodných textoch nasleduje zápisnica z prvej realizovanej vizitácie, konkrétne vo farnosti Valaská. Prvé dva riadky tejto zápisnice obsahujú informáciu, že návšteva vizitátora sa konala dňa 6. októbra roku 1754 („*Visitatio Canonica Parochiae Valaszkensis Anno 1754. die Sexta Octobris peracta*"). Následne sú rovnakým rukopisom spísané vizitačné protokoly farností Lopej, Horná Lehota (fília Lopejskej farnosti), Predajná (s fíliou v osade Jasenie), Dubová (dnes súčasť Nemeckej), Svätý Ondrej (dnes súčasť Brusna), Ľubietová, Poniky, Dúbravica, Zvolenská Ľupča (dnes Slovenská Ľupča), Podkonice (fília farnosti v Ľupči), Selce, Špania Dolina, Staré Hory (fília Španej Doliny) a Radvaň. Zaujímavý je vizitačný protokol mesta Banskej Bystrice (spolu s 13 filiálnymi farnosťami), ktorému predchádza opäť prepis nariadenia Márie Terézie o vykonaní vizitácie z mája roku 1752 a pokyn arcibiskupa Čákiho zo 4. augusta 1754 adresovaný

13 Jozef Baťán (1727 – 1799) sa v roku 1776 stal ostrihomským arcibiskupom. Viac biografických informácií poskytuje *Slovenský biografický slovník, I. zv. (A – D)*. Martin: Matica slovenská, 1986, s. 163 (heslo Jozef Baťán).

14 Môj predpoklad potvrdzuje aj informácia z diela staršieho bystrického historika Emila Jurkoviča (Jurkovicha) *Dejiny kráľovského mesta Banská Bystrica*, že „podklady pre Baťániho vizitáciu v rokoch 1754 – 1755 pripravil jezuitský rektor Wolfgang Ebenhöch." JURKOVIČ, Emil: *Dejiny kráľovského mesta Banská Bystrica.* Banská Bystrica : Občianske združenie Pribicer, 2005, s. 123.

vizitátorovi Baťánovi. Tento vizitačný protokol má formu dotazníka s otázkami a odpoveďami, ktoré pravdepodobne spísal jezuita a zároveň rektor tamojšieho kolégia Wolfgang Ebenhöch. Banskobystrický vizitačný protokol je datovaný k 1. decembru roku 1755. K textu protokolu je následne pripojený aj inventár cenností Kostola Nanebovstúpenia Panny Márie (17 strán) a tiež zoznam kníh farskej knižnice (na dvoch stranách, zoznam obsahuje 142 titulov). Oba pripojené dokumenty sú však písané inou pisárskou rukou ako samotný vizitačný protokol z Banskej Bystrice. Nasleduje protokol z kanonickej vizitácie farnosti Brezno (z roku 1754), ktorý je opäť písaný inou pisárskou rukou. Zaujímavé je, že tento protokol je ako jediný signovaný vizitátorom, bratislavským prepoštom Jozefom Baťánom, a opatrený jeho pečaťou. Potvrdenie pravosti je datované v Bratislave dňa 12. júna 1757 (!). Po štvorstranovom elenchu farností Horného dištriktu arcidiakonátu nasledujú vizitačné protokoly Dolného dištriktu Zvolenskej arcidiakonátu. Konkrétne ide o farnosti: Očová, Hrochoť, Detva, Veľká Slatina (dnes Zvolenská Slatina), Dobrá Niva, Krupina, Babiná, Zvolen, Tŕnie, Budča, Bacúrov, Ostrá Lúka, Badín, Sielnica, Hájniky, Mičiná a Čerín. Za posledným protokolom je pripojený stručný súhrn týchto farností v samostatnom elenchu. V záverečnej časti prameňa sú ešte pripojené dva epištolárne dokumenty – kópia listu vizitárora Jozefa Baťána banskobystrickej komore a originál jej odpovede zo dňa 17. októbra 1754. Posledným textom v súbore je index lokalít, v ktorých sa konali vizitácie. Aj tieto texty sú písané rôznymi pisárskymi rukami a pravdepodobne boli na popud vizitátora zozbierané a zviazané do knižnej podoby. Predpokladám, že čistopisná verzia finálnych protokolov jednotlivých farností Zvolenského arcidiakonátu mala slúžiť potrebám svojho vizitátora a jeho domovskej cirkevnej inštitúcie, ktorou bola v roku 1756 už Bratislavská kapitula (Baťán bol jej prepoštom). Či existujú aj iné, možno originálne protokoly z kanonických vizitácií Zvolenského arcidiakonátu z rokov 1754 a 1755, opatrené pečaťami vizitátora (a možno aj zemepánov s patronátnym právom), sa mi nepodarilo zistiť.[15]

Keďže tento historický prameň som počas svojej prvej návštevy banskobystrického diecézneho archívu objavil aj v zdigitalizovanej podobe, rozhodol som sa na mieste vybrať si ho na spracovanie v projekte SKRIPTOR. Digitalizovanú podobu vizitačného protokolu som získal od vedúceho pracovníka diecézneho archívu Mgr. Petra Kulana v septembri roku 2020. K dispozícii som dostal naskenovanú verziu celého prameňa, ktorá pozostáva zo 184 snímok vo formáte JPEG. V ďalšom texte budem na tieto snímky odkazovať termínom „digitalizáty“ (skratkou: „dig. č.“). Snímky či digitalizáty mali rozlíšenie 300 dpi. Celková veľkosť všetkých digitalizátov mierne presiahla jeden gigabajt dát. V tejto originálnej digitálnej podobe, teda bez akýchkoľvek dodatočných úprav, som prameň nahral do platformy Transkribus a pripravil ho tak na použitie pri vytváraní automatickej transkripcie.

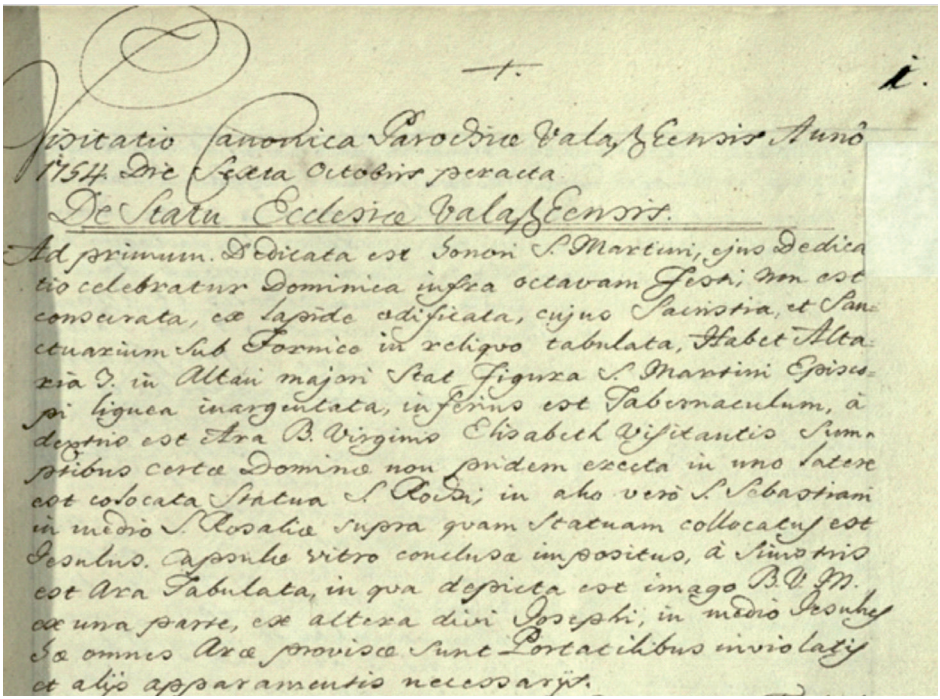
PRÍPRAVA MODELOV AUTOMATICKEJ TRANSKRIPCIE ZO SÚBORU VIZITAČNÝCH PROTOKOLOV ZVOLENSKÉHO ARCIDIAKONÁTU

Po vytvorení vlastného konta v platforme *Transkribus* som v jeho repozitári vytvoril priečinok s názvom „BBBDA – Kanonická vizitácia – CV8“ a nahral som do neho všetkých 184 parciálnych digitalizátov z vizitačných protokolov Zvolenského arcidiakonátu

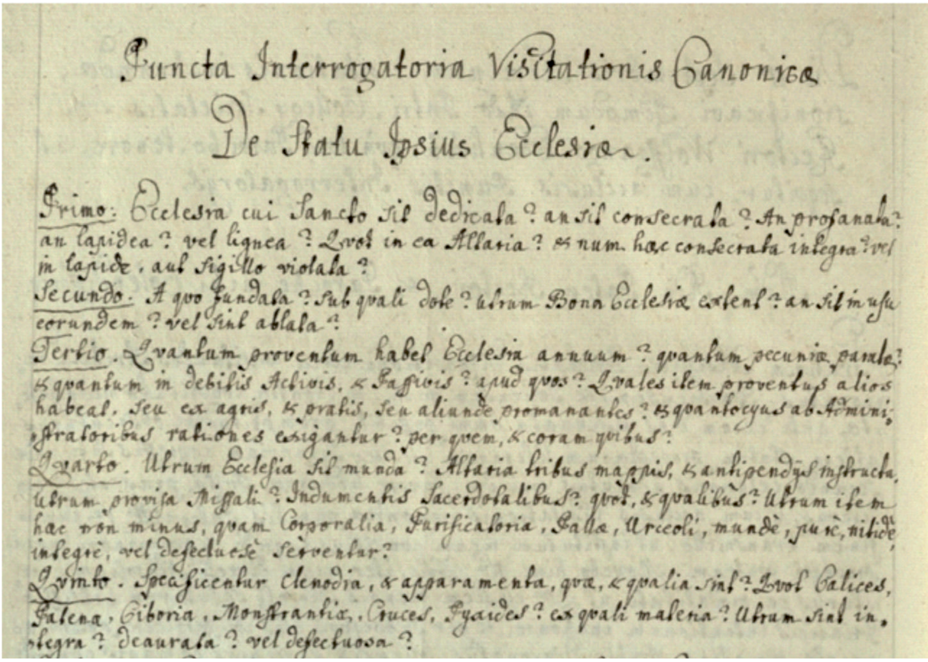
15 Odpoveď na túto otázku nepodáva ani štúdia krupinského historika M. Lukáča (LUKÁČ, Miroslav: K informačnej hodnote katalógu vizitačných protokolov Ostrihomského arcibiskupstva (16. – 20. storočie). In: *Acta historica Neosoliensia*, vol. XXII, č. 1, 2019, s. 45 – 59).

z rokov 1754 – 1756. Na piatich digitalizátoch sa nenachádzal žiadny text, zvyšných 179 obsahovalo zvyčajne dve strany textu, čiže textový rozsah prameňa prekračoval 350 strán. Väčšinou bol text písaný v podobe po sebe nasledujúcich riadkov, ktoré siahali od jedného okraja k druhému, výnimočne bol text rozdelený do dvoch stĺpcov (tak to bolo pri vizitačnom protokole z Banskej Bystrice, dig. č. 72 – 74 a 77 – 79 a č. 91). Ojedinele sa v textoch protokolov vyskytovali jednoduché tabuľky s dvoma či troma stĺpcami – najmä pri enumerácii rôznych príjmov farnosti. Záznamy boli písané vyspelou humanistickou kurzívou v latinčine s použitím početných skratiek, kontrakcií, ligatúr a špecifických znakov (najmä pri uvádzaní dobových peňažných jednotiek – zlatých/florénov a denárov).

Istým hendikepom mnou vybraného prameňa pri vytváraní modelu či viacerých modelov na automatickú transkripciu v platforme *Transkribus* boli rôzne typy rukopisov pri jednotlivých vizitačných protokoloch. Niektoré rozsiahlejšie časti súboru vizitačných protokolov sú písané jednou pisárskou rukou, napríklad 59 digitalizátov od č. 9 po č. 67, prípadne digitalizáty č. 67 – 82 (v rozsahu 16 dvojstrán) alebo aj digitalizáty č. 101 – 173 (v rozsahu 73 dvojstrán). Bolo potrebné sa hneď na začiatku spracovania rukopisného prameňa rozhodnúť, s ktorými časťami budem pracovať pri vytváraní modelov automatickej transkripcie. Napokon som sa rozhodol vybrať si úvodné strany zo všetkých troch dominantne zastúpených rukopisov a vytvárať z nich cvičné modely na trénovanie automatického rozpoznávania a prepisu textov. Na vytvorenie predstavy o podobe rukopisov prezentujem ich podobu s priradením písmen A, B a C.



Obrázok 30 Podoba rukopisu A; výrez z digitalizátu č. 9 (protokol z kanonickej vizitácie farnosti Valaská).



Puncta Interrogatoria Visitationis Canonicae

De Statu Ipsius Ecclesiae.

Primo: Ecclesia cui Sancto sit dedicata? an sit consecrata? An profanata? an lapidea? vel lignea? Quot in ea Altaria? et num. hae consecrata integra? vel in lapide, aut sigillo violata?

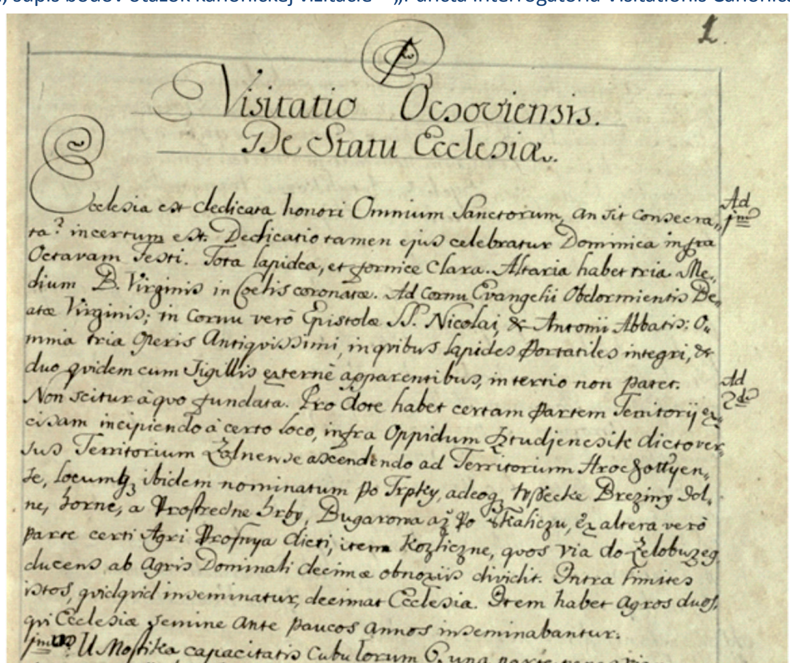
Secundo. A quo fundata? sub quali dote? utrum Bona Ecclesiae extent? an sit in usu eorundem? vel sint ablata?

Tertio. Quantum proventum habet Ecclesia annuum? quantum pecuniae paratae? et quantum in debitis Activis, et Passivis? apud quos? Quales item proventus alios habeat, seu ex agris, et pratis, seu aliunde promanantes? et quo utcunque ab Administratoribus rationes exigantur? per quem, et coram quibus?

Quarto. Utrum Ecclesia sit munda? Altaria tribus mappis, et antipendiis instructa, utrum provisa Missali? Indumentis Sacerdotalibus? quot, et qualibus? Utrum item hac non minus, quam Corporalia, Purificatoria, Palla, Urceoli, munde, pure, nitide integre, vel defectuose serventur?

Quinto. Specificentur Clenodia, et apparamenta, quae, et qualia sint? Quot Calices, Patenae, Ciboria, Monstrantiae, Cruces, Pyxides? ex quali materia? Utrum sint integra? deaurata? vel defectuosa?

Obrázok 31 Podoba rukopisu B; výrez z digitalizátu č. 69 (protokol z kanonickej vizitácie farnosti Banská Bystrica, súpis bodov otázok kanonickej vizitácie – „Puncta Interrogatoria Visitationis Canonicae“).



1.

Visitatio Ocsoviensis.

De Statu Ecclesiae.

Ecclesia est dedicata honori Omnium Sanctorum, an sit consecrata? incertum est. Dedicatio tamen ejus celebratur Dominica infra Octavam Festi. Tota lapidea, et fornice clara. Altaria habet tria. Medium B. Virginis in Coelis coronatae. Ad Cornu Evangelii Obdormientis Beatae Virginis; in Cornu vero Epistolae SS. Nicolai et Antonii Abbatis: Omnia tria Operis Antiquissimi, in quibus lapides Portatiles integri, et duo quidem cum Sigillis externe apparentibus, in tertio non paret. Ad 1mum

Non scitur a quo fundata. Pro dote habet certam Partem Territorii ejusdam incipiendo a certo loco, infra Oppidum Hrudjenesik dicto versus Territorium Zolnense ascendendo ad Territorium Hroc Sottyense, locumque ibidem nominatum po Trpky, adeoque trzecke Breziny dolne, horne, a Prostredne Srby, Dugarovia až po Skaliczu, ex altera vero parte certi Agri Prosnya dicti, item Kozliczne, quos via do Zelobuzeg ducens ab agris Dominali decimae obnoxiis dividit. Intra limites istos, quidquid inseminatur, decimat Ecclesia. Item habet agros duos, qui Ecclesiae semine ante paucos annos inseminabantur. Ad 2dum

1mo. U Mostika capacitatis Cubulorum 6, una parte penes viam, quae

Obrázok 32 Podoba rukopisu C; výrez z digitalizátu č. 101 (protokol z kanonickej vizitácie farnosti Očová).

Po výbere digitalizátov, s ktorými som plánoval pracovať, bolo na prvej porade riešiteľov projektu rozhodnuté, že so segmentáciou textov vybraných na spracovanie pomôžu pracovníčky Univerzitnej knižnice UMB Mgr. Michaela Mikušková a Mgr. Lucia Nižníková, ktoré sú taktiež riešiteľkami projektu SKRIPTOR. Obe kolegyne začali na moju výzvu segmentovať texty vizitačných protokolov písané rukopisom typu A a rukopisom typu B. V prvom prípade začali od úvodného textu (dig. č. 4) až po dig. č. 10, v druhom prípade segmentovali text od dig. č. 67 (pravá strana dvojstrany) po dig. č. 72 (teda spolu šesť dvojstrán). Kolegyne využili možnosť automatickej segmentácie do blokov textu a rozpoznania jednotlivých riadkov, nakoľko je text oboch rukopisov úhľadne písaný v rovnobežných a približne rovnako dlhých riadkoch. V prípade potreby, ak automaticky vložené orámovanie bloku textu „odseklo" nejaké znaky na začiatku alebo konci riadkov, bola hranica bloku ručne posunutá tak, aby blok textu obsiahol kompletný zápis. Výnimočne bola z automaticky segmentovaného textového bloku odstránená časť v podobe jednoduchej tabuľky (najčastejšie s údajmi o príjmoch kňaza či farnosti), ktorej segmentácia sa zdala byť v úvodnej fáze pre riešiteľov náročnejšia (napríklad tabuľka na ľavej strane dig. č. 11 alebo číselné údaje v tabuľke na ľavej strane dig. č. 12). Oprava postupnosti či poradia riadkov takmer nebola nutná, pretože automatická segmentácia bloku textu na jednotlivých stranách ich zoradila správne. Kolegyne z Univerzitnej knižnice UMB spočiatku realizovali aj detailnú segmentáciu riadkovej oblasti (Line region), pomocou ktorej je identifikovaný každý jeden znak v rukopisnom texte daného riadku, ale neskôr bolo od tejto činnosti upustené, nakoľko tento typ segmentácie nemá pre proces automatickej transkripcie bezprostredný význam.

Automatická transkripcia textov v platforme *Transkribus* spočíva v grafickom rozpoznávaní jednotlivých znakov rukopisu vytvoreným softvérovým nástrojom, tzv. modelom. Na vytvorenie takéhoto modelu je potrebné k segmentovanému textu priradiť aj čo najpresnejší prepis. Na vytvorenie modelu je teda potrebné prepísať niekoľko strán textu vybraného prameňa, na ktorých sa softvérový nástroj naučí rozpoznať obrazovú predlohu jednotlivých znakov, aby ich neskôr mohol automaticky správne prečítať a prepísať. Na dosiahnutie čo najlepších výsledkov automatickej transkripcie sa v manuáloch platformy *Transkribus* odporúča prepísať pre potreby cvičného modelu od 5 000 do 15 000 slov zo spracovávaného dokumentu. Keďže však mnou vybrané textové pasáže s tromi typmi rukopisov majú rozsah len pár desiatok strán (konkrétne: rukopis A má celkový rozsah 116 strán, rukopis B má rozsah 29 strán a rukopis C má rozsah 145 strán), rozhodol som sa vytvárať cvičné modely na menšom rozsahu prepísaných textov (zhruba na jednej pätine až štvrtine ich stranového rozsahu). Inšpiroval som sa metodikou vytvorenou študentom magisterského stupňa štúdia Martinom Katreniakom, ktorý pod vedením Ota Tomečka, člena Katedry histórie FF UMB, napísal diplomovú prácu zameranú na uplatnenie automatického prepisu textov kanonických vizitácií z 80. rokov 18. storočia v platforme *Transkribus*.[16] Na základe jeho metodiky a výsledkov, s ktorými som sa zoznámil na jar roku 2022, som aj ja napokon vytváral cvičné modely s menším počtom slov (cca 5 000 – 7 000).

16 KATRENIAK, Martin: *Automatická transkripcia rukopisných historických textov na príklade vybraných kanonických vizitácií* [Diplomová práca]. Školiteľ O. Tomeček, Banská Bystrica : Univerzita Mateja Bela, 2022, 79 s.

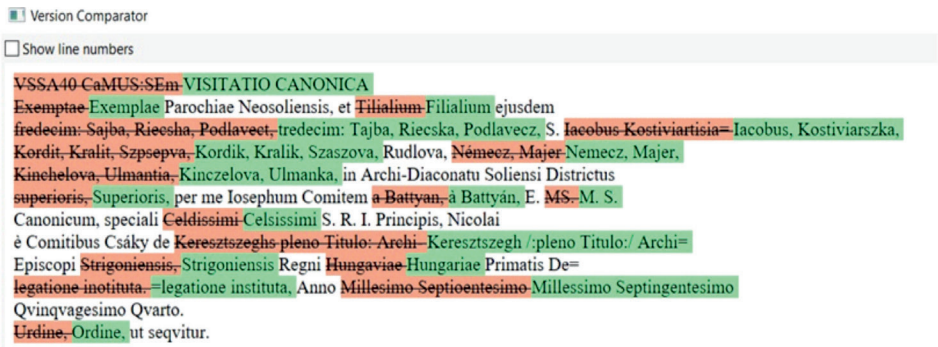
Kým kolegyne z Univerzitnej knižnice UMB v priebehu prvej polovice roku 2021 pracovali nad segmentáciou vybraných pasáží z dvoch prvých rukopisov vizitačných protokolov, realizoval som zatiaľ transkripciu týchto textov. V prípade prvých strán protokolov písaných rukopisom A mi výrazne pomohol kolega Pavol Maliniak, ktorý dal zhruba 20 strán z vizitačných protokolov farnosti Valaská, Lopej a Horná Lehota prepísať (transkribovať) svojim študentom v rámci kurzu novovekej paleografie. Prepisy študentov, odovzdané v textových súboroch WORD, som následne porovnávaním s textom originálneho prameňa skorigoval a postupne vkladal do platformy *Transkribus* metódou jednoduchého kopírovania jednotlivých riadkov. V začiatočnej fáze riešenia projektu som prepísal cca desať strán z vizitačného protokolu mesta Banská Bystrica (rukopis B), ktorého segmentáciu taktiež postupne pripravili kolegyne z Univerzitnej knižnice UMB.

K procesu transkripcie môžem uviesť niekoľko technicko-metodických detailov: snažil som sa transkribovať vybrané pasáže vizitačných protokolov čo najvernejšie k ich textovej predlohe. Ak boli v slovách použité znaky veľkých písmen, často na začiatku slov (ako „Parochus“, „Parochia“ alebo „Ecclesia“), ponechával som ich v tejto podobe aj v prepise. Skratky koncoviek a kontrakcie (najmä pri tituloch, pri substantívach a častých zámenách) som nerozpisoval. Rovnako som verne prepisoval aj písanie skratiek a znakov v hornom indexe. Špeciálne znaky humanistickej kurzívy používanej v barokovom období v strednej Európe, ako napríklad znak „ſ“, som prepisoval ako „s“. Často používané ligatúry „æ“ a „œ“ som prepísal v podobe „ae“ a „oe“. Zachoval som aj prepis znaku „ÿ“. Najväčším problémom bolo vybrať jednotnú formu prepisu znakov, ktoré sa odlišovali svojou veľkosťou. Bolo pomerne náročné rozhodnúť, či išlo o veľkú alebo malú formu litery (tento problém sa týkal najmä znakov C/c, O/o, P/p, S/s a V/v), preto sa v prepisoch originálneho textu objavuje rovnaké slovo niekedy s veľkým začiatočným písmenom, ale aj s malým začiatočným písmenom. Text originálu vizitačných protokolov bol viac-menej dobre čitateľný, ale stretol som sa aj s ťažšie čitateľnými slovami či slovnými spojeniami. Miernym hendikepom v tomto smere bola moja nedostatočná znalosť latinského jazyka, ktorá mi neumožňovala určiť a prepísať slovo podľa významu spojenia či celej vety. V prípade takejto komplikácie som sa snažil vyhľadať možnú správnu podobu nečitateľného slova buď vyhľadávaním správnej podoby na internete alebo komparáciou prepisovanej časti textu s textami publikovaných edícií originálnych vizitačných protokolov (používal som najmä edíciu *Kanonické vizitácie Dunajeckého dekanátu v Spišskom biskupstve z roku 1832*, ktorá síce obsahuje časovo mladšie vizitačné protokoly, ale mnohé formulácie boli zhodné s tými z vizitačných protokolov Zvolenského arcidiakonátu z rokov 1754 – 1756).[17]

Až po vyše roku od riešenia projektu som začal s automatickou segmentáciou tretieho rukopisu (typ C), ktorý sa začína digitalizátom č. 101 (vizitácia farnosti Očová). Text som segmentoval podobným spôsobom ako pri prvých dvoch rukopisných výberoch, pričom som si dovolil vypustiť z blokov segmentovaných textov niekoľko tabuliek s číselnými údajmi, prípadne aj poznámky na margu strán (najmä predložku „ad“ s príslušným číslom: „Ad 1^{mum}“, „Ad 2^{dum}“, „Ad 3^{tium}“ a pod.). Marginálne poznámky som vypustil najmä z toho dôvodu, že presahovali bloky textov jednotlivých úhľadne písaných riadkov. Vyhol som sa tak nutnosti rozširovať textové pole a zložito korigovať správnu postupnosť riadkov so vsunutými marginálnymi poznámkami.

17 ŠIMONČIČ, Jozef – KARABOVÁ, Katarína (eds.): *Kanonické vizitácie Dunajeckého dekanátu v Spišskom biskupstve z roku 1832.* Kraków : Towarzystwo Słowaków w Polsce, 2015. 848 s.

Keď už som mal vytvorené v platforme *Transkribus* prvé prepisy textov kanonických vizitácií z prvého a druhého rukopisu, ešte raz som si ich zbežne skontroloval a prepísané strany som označil ako vzorky „Ground Truth" (teda finálna verzia prepisu). Takto vytvorené cvičné súbory textov som následne použil na vytvorenie prvých dvoch modelov s použitím aplikácie automatického prepisu „HTR+". Na vytváranie prvého modelu som použil relatívne dobre čitateľný prepis strán druhého rukopisu (rukopis typu B), teda časť protokolu kanonickej vizitácie mesta Banská Bystrica. Na vytvorenie cvičného modelu som do cvičného súboru vložil dvanásť prepísaných strán (dig. č. 69 – 74) a do overovacieho súboru len dve strany (dig. č. 67 a 68, obsahujú len po jednej strane písaného textu). Na prepísaných stranách bolo spolu 4 701 slov. Model som nechal trénovať na 50 opakovaniach (epochách). Výsledok chybovosti v znakoch (CER) prvého modelu bol 0,24 % v cvičnom súbore a 6,99 % v overovacom súbore. Pri detailnejšom výpočte presnosti v porovnaní môjho prepisu a automatického prepisu bola dosiahnutá hodnota chybovosti v znakoch (CER) 4,59 % a hodnota chybovosti v slovách (WER) 23,41 %.


Version Comparator

Show line numbers

~~VSSA40 CaMUS:SEm~~ VISITATIO CANONICA
~~Exemptae~~ Exemplae Parochiae Neosoliensis, et ~~Filialium~~ Filialium ejusdem
~~fredecim: Sajba, Riecsha, Podlavect,~~ tredecim: Tajba, Riecska, Podlavecz, S. ~~Iacobus Kostiviartisia=~~ Iacobus, Kostiviarszka,
~~Kordit, Kralit, Szpsepva,~~ Kordik, Kralik, Szaszova, Rudlova, ~~Némecz, Majer~~ Nemecz, Majer,
~~Kinchelova, Ulmantia,~~ Kinczelova, Ulmanka, in Archi-Diaconatu Soliensi Districtus
~~superioris,~~ Superioris, per me Iosephum Comitem ~~a Battyan,~~ à Battyán, E. ~~MS.~~ M. S.
Canonicum, speciali ~~Celdissimi~~ Celsissimi S. R. I. Principis, Nicolai
è Comitibus Csáky de ~~Keresztszeghs pleno Titulo: Archi~~ Keresztszegh /:pleno Titulo:/ Archi=
Episcopi ~~Strigoniensis,~~ Strigoniensis Regni ~~Hungaviae~~ Hungariae Primatis De=
~~legatione inotituta,~~ =legatione instituta, Anno ~~Millesimo Septioentesimo~~ Millessimo Septingentesimo
Qvinqvagesimo Qvarto.
~~Urdine,~~ Ordine, ut seqvitur.

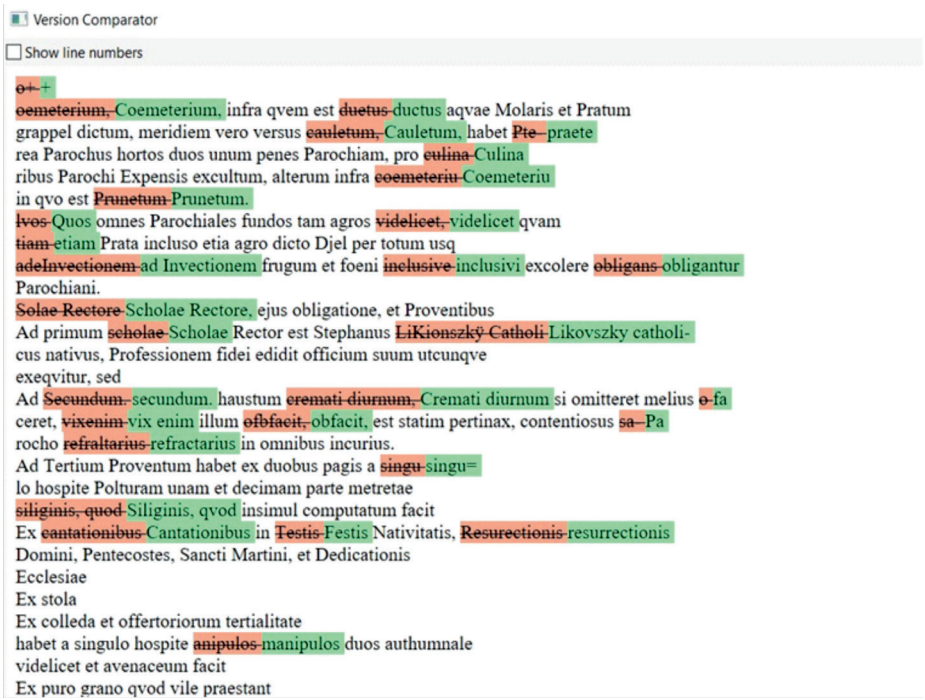
Obrázok 33 Výsledok uplatnenia modelu na overovacom súbore (dig. č. 67).

Vysvetlivka: biele podfarbenie = správny strojový prepis, červené podfarbenie = nesprávny prepis, zelené podfarbenie = môj pôvodný prepis originálu prameňa

Hodnotenie úspešnosti prepisu v modeli: Keďže som v cvičnom súbore použil relatívne malý rozsah textu, ktorý obsahoval len 4 700 slov, výsledok transkripcie v overovacom súbore nebol príliš uspokojivý. Chybovosť sa prejavila najmä v nerozpoznaní veľkých písmen, ktoré boli v cvičnom súbore zastúpené príliš málo, ako aj v zámene niektorých hlások (napríklad často bola zamenená hláska „l" za „t", prípadne opačne, alebo aj hláska „t" za „r", ak bolo „t" písané v strede slova). Početné chyby však spočívali aj v absencii interpunkčných znamienok a čiarok, ktoré model nedokázal v overovacom súbore správne identifikovať a zapísať.

Druhý cvičný model som vytváral z textu prvého rukopisu (rukopis typu A) vizitačných protokolov farností Valaská a Lopej. Týmto textom ešte predchádzal prepis listu arcibiskupa Čákiho, ktorý však bol písaný inou písarskou rukou. Na vytvorenie modelu som do cvičného súboru vložil 20 prepísaných strán (dig. č. 6 – 16) a do overovacieho súboru len dve dvojstrany (dig. č. 17 a 18). V prepísaných stranách cvičného súboru bolo 5 482 slov, v overovacom súbore 1 240 slov. Model som nechal trénovať na 50 opakovaniach (epochách). Výsledok chybovosti v znakoch (CER) druhého modelu

bol 0,51 % v cvičnom súbore a 7,05 % v overovacom súbore. Pri výpočte presnosti v porovnaní môjho prepisu a automatického prepisu bola dosiahnutá hodnota CER 6,56 %. Chybovosť v slovách (WER) však dosahovala pri tomto modeli takmer 30 %.


Version Comparator

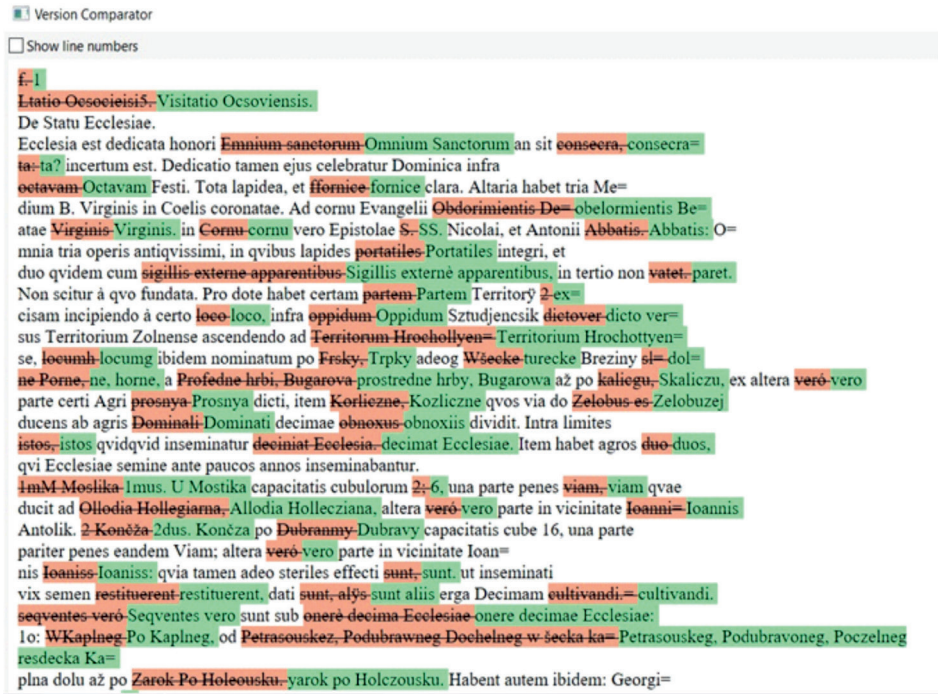
Show line numbers

o+ +
oemeterium, Coemeterium, infra qvem est ductus ductus aqvae Molaris et Pratum
grappel dictum, meridiem vero versus cauletum, Cauletum, habet Pte praete
rea Parochus hortos duos unum penes Parochiam, pro culina Culina
ribus Parochi Expensis excultum, alterum infra coemeteriu Coemeteriu
in qvo est Prunetum Prunetum.
lvos Quos omnes Parochiales fundos tam agros videlicet, videlicet qvam
tiam etiam Prata incluso etia agro dicto Djel per totum usq
adeInvectionem ad Invectionem frugum et foeni inclusive inclusivi excolere obligans obligantur
Parochiani.
Solae Rectore Scholae Rectore, ejus obligatione, et Proventibus
Ad primum scholae Scholae Rector est Stephanus LiKionszkÿ Catholi Likovszky catholi-
cus nativus, Professionem fidei edidit officium suum utcunqve
exeqvitur, sed
Ad Secundum. secundum. haustum eremati diurnum, Cremati diurnum si omitteret melius o fa
ceret, vixenim vix enim illum ofbfacit, obfacit, est statim pertinax, contentiosus sa Pa
rocho refraltarius refractarius in omnibus incurius.
Ad Tertium Proventum habet ex duobus pagis a singu singu=
lo hospite Polturam unam et decimam parte metretae
siliginis, quod Siliginis, qvod insimul computatum facit
Ex cantationibus Cantationibus in Testis Festis Nativitatis, Resurectionis resurrectionis
Domini, Pentecostes, Sancti Martini, et Dedicationis
Ecclesiae
Ex stola
Ex colleda et offertoriorum tertialitate
habet a singulo hospite anipulos manipulos duos authumnale
videlicet et avenaceum facit
Ex puro grano qvod vile praestant

Obrázok 34 Výsledok uplatnenia 2. modelu na overovacom súbore (dig. č. 18).

Hodnotenie úspešnosti prepisu v modeli: Častou chybou v overovacom súbore boli opäť chybne prepísané písmená, ktoré sú si graficky podobné, ako napr. c/e, t/r alebo T/F. Opakovane sa vyskytol problém s rozpoznaním a správnym prepísaním malých a veľkých písmen (c/C, p/P, s/S), prípadne neboli správne rozpoznané podobne vyzerajúce veľké písmená (napr. Testis – správne Festis). Model tiež niekedy spájal slová (teda ignoroval medzery). Výsledná podoba prepisu v overovacom súbore sa mi však javila ako lepšia, ako bol výsledok v prípade prvého modelu.

Napokon tretí cvičný model som vytvoril na základe prepisu rukopisu protokolu kanonickej vizitácie farností Očová a Hrochoť (desať strán textu, dig. č. 103 – 107). Na prepísaných stranách civčného súboru bolo iba 3 285 slov, pretože na týchto stranách sa nachádzali aj dlhšie zoznamy mien farníkov, ktorých zápis zabral len 1/3 alebo 1/2 riadku. Model som ale tentoraz nechal trénovať na 100 opakovaniach (epochách), ale zistil som, že dvojnásobné navýšenie počtu opakovaní neprispelo k lepšiemu výsledku. Výsledok chybovosti v znakoch (CER) tretieho modelu bol 0,12 % v cvičnom súbore a 6,97 % v overovacom súbore. Pri výpočte presnosti v porovnaní môjho prepisu a automatického prepisu bola dosiahnutá hodnota chybovosti v znakoch (CER) 7,24 %, ale chybovosť v slovách (WER) dosahovala viac ako 27 %.


Version Comparator

Show line numbers

f. 1
Ltatio Ocsocieisi5. Visitatio Ocsoviensis.
De Statu Ecclesiae.
Ecclesia est dedicata honori Emnium sanctorum Omnium Sanctorum an sit consecra, consecra=
ta; ta? incertum est. Dedicatio tamen ejus celebratur Dominica infra
octavam Octavam Festi. Tota lapidea, et ffornice fornice clara. Altaria habet tria Me=
dium B. Virginis in Coelis coronatae. Ad cornu Evangelii Obdorimientis De= obelormientis Be=
atae Virginis Virginis, in Cornu cornu vero Epistolae S. SS. Nicolai, et Antonii Abbatis. Abbatis: O=
mnia tria operis antiqvissimi, in qvibus lapides portatiles Portatiles integri, et
duo qvidem cum sigillis externe apparentibus Sigillis externe apparentibus, in tertio non vatet. paret.
Non scitur à qvo fundata. Pro dote habet certam partem Partem Territorÿ 2 ex=
cisam incipiendo à certo loco loco, infra oppidum Oppidum Sztudjencsik dictover dicto ver=
sus Territorium Zolnense ascendendo ad Territorum Hrochollyen= Territorium Hrochottyen=
se, locumh locumg ibidem nominatum po Frsky, Trpky adeog Wšecke turecke Breziny sl= dol=
ne Porne, ne, horne, a Profedne hrbi, Bugarova prostredne hrby, Bugarowa až po kaliegu, Skaliczu, ex altera veró vero
parte certi Agri prosnya Prosnya dicti, item Korliezne, Kozliczne qvos via do Zelobus es Zelobuzej
ducens ab agris Dominali Dominati decimae obnoxus obnoxiis dividit. Intra limites
istos, istos qvidqvid inseminatur deciniat Ecclesia. decimat Ecclesiae. Item habet agros duo duos,
qvi Ecclesiae semine ante paucos annos inseminabantur.
1mM Moslika 1mus. U Mostika capacitatis cubulorum 2, 6, una parte penes viam, viam qvae
ducit ad Ollodia Hollegiarna, Allodia Hollecziana, altera veró vero parte in vicinitate Ioanni= Ioannis
Antolik. 2 Končža 2dus. Končza po Dubranmy Dubravy capacitatis cube 16, una parte
pariter penes eandem Viam; altera veró vero parte in vicinitate Ioan=
nis Ioaniss Ioaniss: qvia tamen adeo steriles effecti sunt, sunt. ut inseminati
vix semen restituerent restituerent, dati sunt, alÿs sunt aliis erga Decimam cultivandi. cultivandi.
seqventes veró Seqventes vero sunt sub onerè decima Ecclesiae onere decimae Ecclesiae:
1o: WKaplneg Po Kaplneg, od Petrasouskez, Podubrawneg Dochelneg w šecka ka= Petrasouskeg, Podubravoneg, Poczelneg
resdecka Ka=
plna dolu až po Zarok Po Holeousku. yarok po Holczousku. Habent autem ibidem: Georgi=

Obrázok 35 Výsledok uplatnenia 3. modelu na overovacom súbore (dig. č. 101).

Hodnotenie úspešnosti prepisu v modeli: Početné chyby boli opäť najmä v osobných menách a názvoch topografických lokalít, z ktorých väčšina bola vo vizitačnom protokole uvedená v slovenskej podobe. Tentoraz už model správne prepisoval veľké písmená, minimálne boli aj zámeny veľkých a malých písmen pri rovnakých grafémach (ako c/C). Časté chyby sa objavili pri menách a priezviskách sedliakov, ktorí odovzdávali farárovi desiatok, tie boli často deformované – napr. meno „Mathiam Gonda“ model prepísal ako „Mathiam sonda“, namiesto „Ioannem Gall“ model prepísal „Ioannem Sall“. Znaky, ktoré sa v krátkom cvičnom súbore takmer vôbec nevyskytli – napríklad archaické ž v podobe znaku ʒ alebo graféma w/W, model nevedel prepísať vôbec (meno „Koʒuch“ prepísal ako „Kojuch“, spojenie „Ioannem Wanko“ model prepísal „Ioannemanko“). Relatívne s veľkou presnosťou prepísal model slová písané malými písmenami, ako je viditeľné na obrázku č. 35.

Z výsledkov jednotlivých cvičných modelov, ktoré som vytvoril na relatívne malej vzorke prepísaných textov, bolo zrejmé, že modely s menšou mierou chybovosti potrebujú pravdepodobne väčší objem prepísaného textu v cvičnom súbore, aby sa pri rozpoznávaní jednotlivých grafém stroj „naučil“ správne rozpoznať čo najviac z nich. Z tohto dôvodu som sa rozhodol vylepšiť práve tretí model tým, že do cvičného súboru pridám ďalšie prepísané strany z protokolov.

V novom, vylepšenom modeli na podklade rukopisu typu C som teda vložil do cvičného súboru už 11 dvojstrán (s počtom 7 134 slov). Model som nechal trénovať na 50

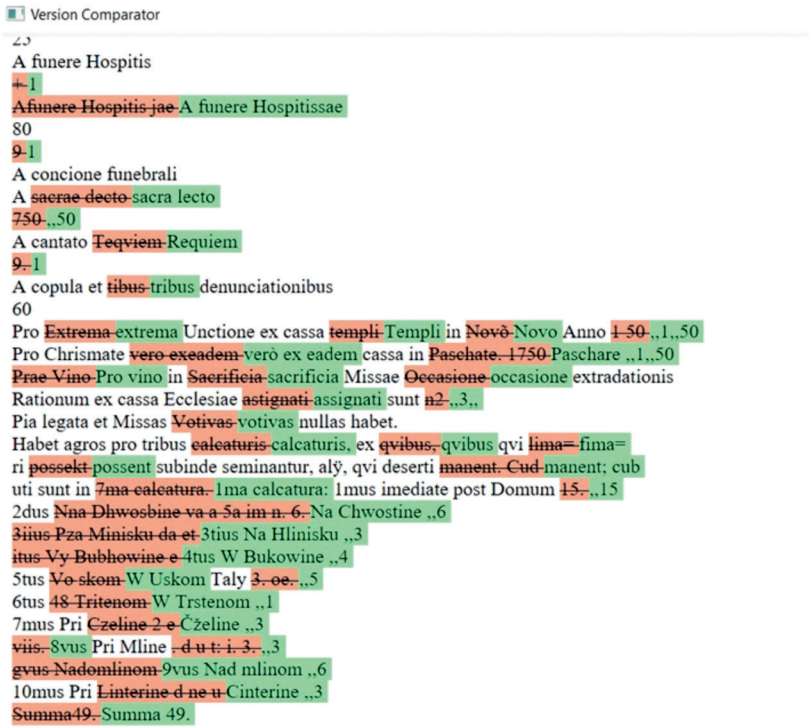
opakovaniach (epochách). Výsledok chybovosti v znakoch (CER) vylepšeného tretieho modelu bol 0,32 % v cvičnom súbore a 6,07 % v overovacom súbore. Pri výpočte presnosti v porovnaní môjho prepisu a automatického prepisu bola dosiahnutá hodnota chybovosti v znakoch (CER) 5,82 % a chybovosť v slovách (WER) dosiahla hodnotu 22,65 %. V oboch parametroch tak došlo k istému zlepšeniu výsledku automatického prepisu overovacieho súboru, o takmer 5 % klesla aj chybovosť v slovách. Zlepšenie automatického prepisu bolo zreteľné aj pri komparatívnom zobrazení strojového prepisu s mojím prepisom originálu prameňa – program sa naučil čítať už aj znaky, ktorých rozpoznávanie mu v predošlej verzii tretieho modelu robilo problémy (napríklad archaickú grafému ȝ). V niektorých prípadoch už vylepšený model správne prečítal aj znak G v osobných menách (napríklad v spojení „Apud Dnam Gruberianam“, ale priezviská „Gonda“ a „Gall“ vylepšený model naďalej prepísal chybne ako „Sonda“ a „Tall“).

K zlepšeniu výsledku automatického prepisu vylepšeného tretieho modelu teda došlo, ale pri zlepšovaní modelu bolo v cvičnom a overovacom súbore použitých celkovo 25 strán textu z vizitačných protokolov rukopisu typu C (jeho celkový rozsah je 145 strán textu, dig. č. 101 – 173). Na prípravu modelu automatického prepisu v platforme *Transkribus* bola teda využitá jedna šestina rozsahu prameňa s identickým rukopisom. Pri tvorbe modelov automatického prepisu stojí za zváženie, aký by mal byť minimálny – ale aj maximálny – počet strán textu (alebo presnejšie slov v tomto textovom súbore) potrebných na vytvorenie funkčného modelu. Najmä pri prameňoch s identickým rukopisom s malým rozsahom (30 – 50 strán) sa mi javí ako málo efektívne spracovať pre potreby vytvorenia modelu takmer polovicu ich rozsahu. Osobne za efektívne považujem využitie zhruba jednej pätiny, prípadne jednej štvrtiny rozsahu prameňa na vytvorenie funkčného modelu na automatický prepis zvyšku rukopisu, pričom by bola dosiahnutá maximálna chybovosť v znakoch v hodnote okolo 6 – 7 %. Na mojom príklade vytvárania malých modelov automatickej transkripcie pri kratších rukopisných prameňoch sa ukazuje, že túto chybovosť dosahujú modely vypracované aj na základe využitia 5 000 – 7 000 slov v cvičnom súbore.

Napokon som sa rozhodol ešte vyskúšať možné vylepšenie druhej verzie tretieho modelu s použitím už vytvoreného modelu, ktorý je v bezplatnej ponuke modelov platformy *Transkribus* (v platforme sú označované ako Base Models, teda základné modely). Na základe odporúčaní kolegov (napríklad O. Tomečka), ako aj na podklade aplikácie jedného zo základných modelov vo vylepšovaní modelov vytváraných študentom Martinom Katreniakom v jeho diplomovej práci, som sa rozhodol použiť model s názvom *NeoLatin_Ravenstein*, ktorý bol vytvorený taktiež na podklade latinských prameňov cirkevnej povahy z obdobia 17. a 18. storočia.[18] Na cvičenie modelu som vybral niekoľko strán z dvoch podobných rukopisov – teda rukopisu typu A a C. Spolu bolo v cvičnom súbore 12 733 slov z 20 digitalizátov. Overovací súbor som vytvoril z desiatich strán textu (teda piatich digitalizátov, dig. č. 110 – 114). Počet opakovaní (epoch) som stanovil na 50. Základný model *NeoLatin_Ravenstein* som aplikoval na môj vylepšený tretí model

18 Model „NeoLatin_Ravenstein“ je založený na prepise výročných listov Litterae Annuae Parochiae Ravensteijn SJ ab Anno 1643 ad Annum 1772, ktoré sa momentálne nachádzajú v Katolíckom dokumentačnom centre v Leuvene v Belgicku. Model obsahuje 64 435 slov na 6 864 riadkoch (115 strán v cvičnom súbore a 8 strán v overovacom súbore). Chybovosť modelu v znakoch predstavuje hodnotu 3,58 % v overovacom súbore a 4,51 % v cvičnom súbore.

(s chybovosťou v znakoch v overovacom súbore na úrovni cca 6 %). Keďže M. Katreniakovi sa pri uplatnení tohto základného modelu podarilo vylepšiť trénované modely v rovine znakovej chybovosti v rozmedzí od 0,48 % po 3,25 %,[19] očakával som aj ja zlepšenie (teda zníženie) miery chybovosti CER v overovacom súbore. Výsledok ma však mierne prekvapil. Výsledok chybovosti v znakoch (CER) kombinovaného modelu bol 1,78 % v cvičnom súbore a 7,06 % v overovacom súbore. Pri výpočte presnosti v porovnaní môjho prepisu a automatického prepisu bola dosiahnutá hodnota chybovosti v znakoch (CER) 7,69 % a chybovosť v slovách (WER) dosiahla hodnotu 28,42 %. Ako som však zistil pri detailnejšej kontrole výsledkov transkripcie týmto kombinovaným modelom, vysoká miera chybovosti v znakoch (CER) – až 18,85 % – bola zaznamenaná pri tretej dvojstrane (dig. č. 112) overovacieho súboru, na ktorej boli v origináli prameňa uvedené topografické lokality farnosti v Detve, písané v slovenskom jazyku. Ich zápis obsahoval špecifické znaky (napríklad: č, ǯ, ch alebo w), ako aj početné interpunkčné znamienka, ktoré kombinovaný model nedokázal správne prečítať.



Obrázok 36 Výsledok uplatnenia kombinovaného modelu na overovacom súbore (dig. č. 112).

Pri ďalších štyroch dvojstranách vložených do overovacieho súboru dosiahla chybovosť v znakoch (CER) hodnoty od 5,18 % po 6,27 %, čo považujem za veľmi dobrý výsledok. Po realizácii opisovaného pokusu o vylepšenie môjho tretieho modelu jeho kombináciou so základným modelom *NeoLatin_Ravenstein* som nateraz prerušil tvorbu ďalších modelov v platforme *Transkribus*.

19 KATRENIAK, Martin: *Automatická transkripcia rukopisných historických...*, s. 60 – 62.

ZÁVER

V projekte SKRIPTOR som si na výskum a experimenty vybral špecifický dobový prameň úradnej povahy – protokoly z kanonických vizitácií farností Zvolenského arcidiakonátu, ktoré vznikli v rozmedzí rokov 1754 – 1756. Istý hendikep pri použití tohto typu prameňa v platforme *Transkribus* predstavuje fakt, že ide o súbor niekoľkých textových dokumentov rozličnej formy a obsahu (listy, hodnotiace správy z vizitácie, dotazníková forma správy, inventárne zoznamy cenností a kníh, sumáre a indexy), a tiež fakt, že tieto dokumenty boli vytvorené rozličnými autormi (pisárskymi rukami). Práve variabilita rukopisov, aj keď sú si do veľkej miery podobné, do istej miery komplikuje vytvorenie takého modelu, ktorý by dokázal tieto texty automaticky transkribovať s čo najväčšou presnosťou a s čo najmenšou chybovosťou. Počas vytvárania modelov vybraného dobového prameňa v platforme *Transkribus* som uplatnil prístup tvorby parciálnych modelov pre tri dominantné rukopisy vizitačných protokolov. Hoci išlo o modely vytvárané na malej vzorke prepísaných textov, napokon ma pomerne nízka miera ich chybovosti pozitívne prekvapila. Pri treťom modeli som sa pokúsil o jeho zlepšenie dvojnásobným navýšením počtu slov v cvičnom súbore, pričom som dosiahol zníženie chybovosti o jedno percento (z približne 7 % na 6 %). Miera chybovosti modelov bola ovplyvnená najmä nízkym zastúpením špecifických znakov, pri ktorých mali vytvorené modely problém naučiť sa ich správne rozpoznať a následne prepísať. Predpokladám, že postupným vylepšovaním vytvorených modelov dokážem postupne znižovať mieru chybovosti pri automatickom prepise overovacích súborov textu prameňa, ale aj doteraz neprepísaných a nesegmentovaných častí vizitačných protokolov, ktoré v projekte SKRIPTOR spracovávam. V budúcnosti chcem ďalej vylepšovať už vytvorené modely na automatickú transkripciu a možno aj pribrať na spracovanie ďalší, nový súbor protokolov z kanonických vizitácií, pravdepodobne z rovnakého časového obdobia (z druhej polovice 18. storočia). Výsledky automatickej transkripcie častí týchto prameňov v platforme *Transkribus* chcem – po konzultácii s kolegami z pracoviska – ponúknuť na prípadné publikovanie v kritickej, možno bilingválnej (latinsko-slovenskej) edícii.

V záverečnom zhodnotení výsledkov mojej aktivity pri riešení úloh projektu SKRIPTOR môžem konštatovať, že spoznanie možností praktického využitia platformy *Transkribus*, konkrétne jej softvérových nástrojov vytváraných na automatickú transkripciu historických písomných prameňov, bolo a je pre mňa obohacujúce. Počiatočné problémy s pochopením jednotlivých technických krokov a metód pri príprave cvičných modelov automatickej transkripcie sa podarilo prekonať a dosiahnuté pozitívne výsledky sú pre mňa stimulom na zdokonaľovanie a následné praktické použitie nástrojov automatickej transkripcie vybraných historických prameňov.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

BAŤÁN, Jozef. In: *Slovenský biografický slovník*, I. zv. (*A – D*). Martin : Matica slovenská, 1986, 544 s.

BOŠANSKÝ, Martin: *Kánonické vizitácie Nitrianskej diecézy do roku 1831.* Nitra : Kňazský seminár sv. Gorazda, 2018. 220 s. ISBN 978-80-89481-47-7.

JURKOVIČ, Emil: *Dejiny kráľovského mesta Banská Bystrica*. Banská Bystrica : Občianske združenie Pribicer, 2005. 552 s. ISBN 80-969366-2-X.

KATRENIAK, Martin: *Automatická transkripcia rukopisných historických textov na príklade vybraných kanonických vizitácií.* [Diplomová práca.] Banská Bystrica : Filozofická fakulta UMB, 2022. 79 s.

LOPATKOVÁ, Zuzana: Edície zápisníc za kanonických vizitácií z obdobia novoveku. In: KOHÚTOVÁ, Mária, Zuzana LOPATKOVÁ a kol. *Metodologické limity historického prameňa.* Kraków; Trnava : Towarzystwo Słowaków w Polsce; Filozofická fakulta Trnavskej univerzity, 2014, s. 38 – 47. ISBN 978-83-7490-795-8.

LOPATKOVÁ, Zuzana: *Kanonické vizitácie 16. – 18. storočia v slovenských dejinách.* Trnava : Filozofická fakulta Trnavskej univerzity, 2021. 64 s. ISBN 978-80-568-0442-1.

LUKÁČ, Miroslav: K informačnej hodnote katalógu vizitačných protokolov Ostrihomského arcibiskupstva (16. – 20. storočie). In: *Acta historica Neosoliensia,* roč. XXII, č. 1, 2019, s. 45 – 59.

MÚDRA, Darina: *Topografia hudby klasicizmu na Slovensku z pohľadu kanonických vizitácií.* Bratislava : Veda, vydavateľstvo SAV, 2019. 1318 s. ISBN 978-80-224-1742-6.

ŠIMONČIČ, Jozef a Katarína KARABOVÁ: *Kanonické vizitácie Dunajeckého dekanátu v Spišskom biskupstve z roku 1832.* Kraków : Towarzystwo Słowaków w Polsce, 2015. 848 s.

pk [KULAN, Peter]: *Založenie Banskobystrickej diecézy.* Dostupné na: https://bbdieceza.sk/banskobystricka-dieceza/historia-diecezy/zalozenie-banskobystrickej-diecezy/

Diecézny archív. Dostupné na: https://bbdieceza.sk/diecezny-archiv/

KAPITOLA 5

AUTOMATICKÁ TRANSKRIPCIA REAMBULAČNÉHO PROTOKOLU BANSKEJ BYSTRICE Z ROKU 1820

Oto Tomeček
Univerzita Mateja Bela v Banskej Bystrici; Filozofická fakulta; Katedra histórie
E-mail: oto.tomecek@umb.sk

ABSTRAKT

Kapitola opisuje experiment, ktorého cieľom bolo vytvoriť model na automatickú transkripciu rukopisu reambulačného protokolu mesta Banská Bystrica. Dokument zaznamenáva reambuláciu mestských hraníc vykonanú v letných mesiacoch v roku 1820. Analyzovaný banskoštiavnický exemplár protokolu (okrem neho poznáme aj banskobystrický exemplár) predstavuje čistopis s 245 stranami textu, ktorý je napísaný jednou pisárskou rukou novolatinskou kurzívou. Okrem formálneho popisu dokumentu je v úvode zhodnotený aj obsah, historický význam a okolnosti vzniku dokumentu. Druhá časť práce je venovaná jednotlivým krokom práce v systéme *Transkribus*, ktoré predchádzali vytvoreniu modelu automatickej transkripcie. Podrobne opisuje spôsob nasnímania dokumentu, jeho nahrania do systému *Transkribus* a jeho následnú segmentáciu. Na konci tejto časti je predstavený spôsob manuálneho prepisu vybranej časti dokumentu, ktorá predstavuje vzorku na vytvorenie modelu automatickej transkripcie. Hlavná časť práce predstavuje pracovný postup pri vytvorení štyroch konkrétnych modelov určených na automatickú transkripciu dokumentu.

Kľúčové slová: historický rukopis; reambulácia; reambulačný protokol; *Transkribus*; transkripcia; cvičný súbor, overovací súbor, základný model

ABSTRACT

Automatic transcription of the reambulation protocol of Banská Bystrica from the year 1820

The chapter in question describes an experiment whose goal was to create a model for the automatic transcription of the manuscript of the reambulatory protocol of the city of Banská Bystrica. The document records the reambulation of the city limits carried out in the summer months of the year 1820. The analyzed copy of the protocol from Banská Štiavnica (in addition to it, we also know of the copy from Banská Bystrica) represents a pure copy with 245 pages of text, which is written by one scribe's hand in new Latin cursive. In addition to the formal description of

the document, the introduction also evaluates the content, historical significance and circumstances of the creation of the document. The second part of the work discusses the individual steps of work in the *Transkribus* system, which preceded the creation of the automatic transcription model. It describes in detail the method of scanning a document, uploading it to the *Transkribus* system and its subsequent segmentation. At the end of this section, a method of manual transcription of a selected part of the document is presented, which represents a sample for creating an automatic transcription model. The main part of the work presents the work procedure for the creation of four specific models intended for automatic document transcription.

Key words: historical manuscript; reambulation; reambulatory protocol; *Transcribus*; transcription; training set, validation set, base model

ÚVOD

Reambulačný protokol mesta Banská Bystrica z roku 1820 predstavuje podrobný záznam o priebehu obhliadky (reambulácie) mestských hraníc s ich detailným popisom.[1] Obchádzanie hraníc územných celkov (feudálnych panstiev, mestských a dedinských chotárov) patrilo už od stredoveku k dôležitým povinnostiam miestneho úradníckeho aparátu a predstaviteľov obecných samospráv. Vymedzenie chotárov miest a dedín, ako aj hájenie stability ich hraníc bolo základom ekonomickej prosperity mestských a vidieckych komunít. Práve vo vnútri ich chotárov sa sústreďovala prevažná väčšina hospodárskych aktivít miestneho obyvateľstva. Poznať hranice územia, v ktorej komunita žila a realizovala svoje hospodárske aktivity, nebolo z tohto dôvodu len povinnosťou, ale zároveň aj právom. V takomto duchu chápali obhliadky mestského teritória aj mešťania Banskej Bystrice.

Najstaršie informácie o pravidelných obhliadkach hraníc Banskej Bystrice máme od 16. storočia. Ide predovšetkým o písomné zmienky týkajúce sa finančných výdavkov spojených s pohostením počas týchto obhliadok alebo správy obsahujúce urgencie na opravu poškodených hraničných medzníkov.[2] Pravidelné obhliadky mestských hraníc Banskej Bystrice spomína aj barokový učenec Matej Bel, ktorý vo svojom diele upozornil na značný rozsah mestského teritória a dĺžku jeho hraníc. Podľa neho aj jazdcovi na koni trvalo obídenie hraníc mesta po celom obvode takmer tri celé dni.[3]

Podrobná obhliadka hraníc v roku 1820 vznikla v období vrcholiacich sporov medzi zástupcami mesta a miestnou komorskou správou vo veci hospodárskeho využívania rozsiahleho územia banskobystrického teritória. Vzájomné spory medzi mestom

1 BUMBA, Jan: České *katastry od 11. do 21. století*. Praha : Grada, 2007, s. 182. Užšie sa môže chápať reambulácia aj ako oprava a doplnenie starších máp na základe nových meraní.

2 JURKOVICH, Emil: *Dejiny kráľovského mesta Banská Bystrica*. preklad I. Nagy, Banská Bystrica : Pribicer, 2005, s. 194.

3 BEL, Matej: *Zvolenská stolica*. eds. I. Nagy – M. Turóci, Čadca : Kysucké múzeum, 2017, s. 201.

a banskou komorou pritom boli staršieho dáta. Viac-menej pravidelne sa objavovali už od čias vlády cisára Maximiliána II., ktorý vo svojom banskom poriadku vydanom v roku 1573 rezervoval lesy v okolí banských miest na banské účely.[4] Keď v roku 1819 začalo mesto aj komora vyhľadávať a cielene zhromažďovať písomné dokumenty potvrdzujúce ich majetkové práva na tunajšie lesy,[5] bolo jasné, že sa schyľuje k riešeniu sporu prostredníctvom súdneho konania. Hoci súdny proces medzi mestom a komorou, známy ako kompromisný proces, začal sa až v roku 1835,[6] mesto sa nepochybne na jeho konanie pripravovalo dlhší čas. Zbieralo podklady, ktoré mu mohli pomôcť počas súdneho pojednávania. Zdá sa byť preto pravdepodobné, že aj podrobný reambulačný protokol, obsahujúci detailný popis mestských hraníc, vznikol práve v tejto súvislosti.

Reambulačný protokol označený jednoslovným názvom *Metales* bol vyhotovený minimálne v dvoch exemplároch. Prvý exemplár si ponechalo mesto, druhý dostala k dispozícii komorská správa v Banskej Bystrici. Oba exempláre sa v pomerne dobrom stave zachovali až do súčasnosti. Komorský exemplár je dnes uložený vo fonde Banská komora v Banskej Bystrici v banskoštiavnickom banskom archíve.[7] Exemplár patriaci mestu bol uložený v archíve mesta, dnes v Štátnom archíve v Banskej Bystrici.[8] Obsahovo sú oba exempláre takmer úplne totožné, hoci každý z nich bol napísaný inou pisárskou rukou. Rozloženie textu na jednotlivých stranách je aj z tohto dôvodu v oboch dokumentoch rozdielne. Pozorovať možno aj niektoré ďalšie drobné odlišnosti. Iné je napríklad skracovanie slov alebo písanie veľkých a malých písmen. Jediným zásadnejším obsahovým rozdielom medzi nimi je, že banskobystrický exemplár má na rozdiel od banskoštiavnického exempláru aj titulný list s úplným názvom dokumentu. Podľa neho vieme, že jeho názov (*Metalis Reambulatio inter Terrenum L: R: ac Mont: Civitatis Neosoliensis et Omnia ejus vicina Territoria Anno 1820 peracta*) v preklade znamená *Hraničná reambulácia medzi územím slobodného kráľovského a banského mesta Banská Bystrica a všetkými jeho susediacimi teritóriami vykonaná v roku 1820*. Predmetom môjho podrobnejšieho výskumu a experimentu zameraného na automatickú transkripciu sa stal komorský alebo banskoštiavnický exemplár reambulačného protokolu.

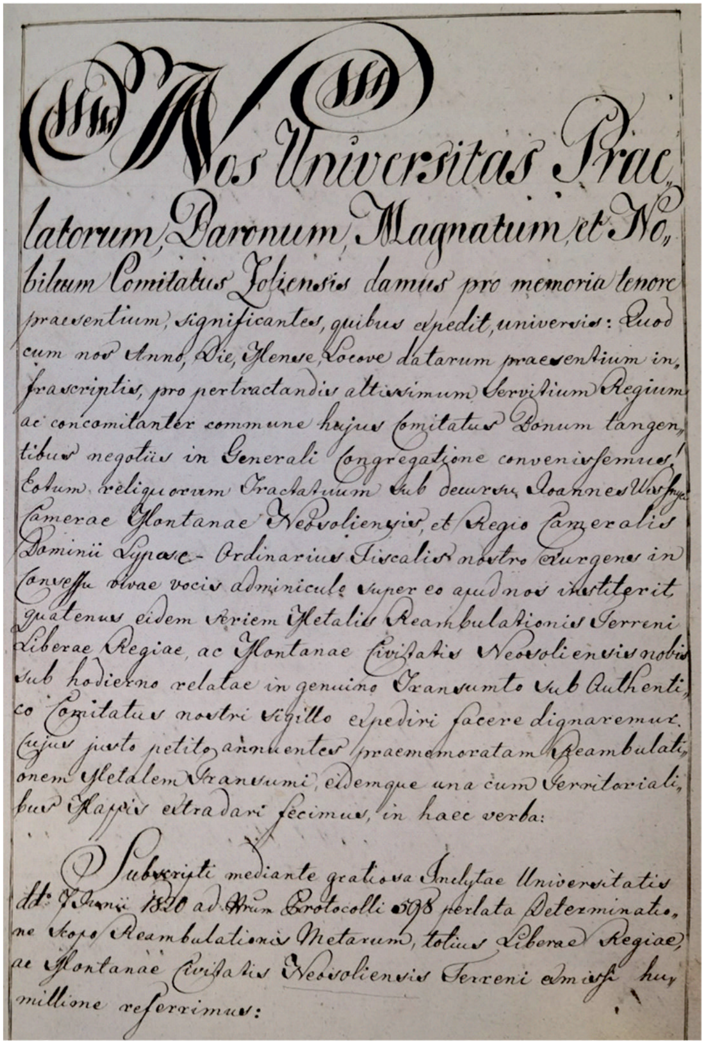
4 TOMEČEK, Oto: *Drevorubači a uhliari v lesoch Banskej Bystrice: k problematike osídlenia horských oblastí banskobystrického chotára do polovice 19. storočia.* Banská Bystrica : Fakulta humanitných vied UMB, 2010, s. 23.

5 JURKOVICH, Emil: *Dejiny kráľovského mesta Banská Bystrica*, s. 179.

6 Rozhodnutie o konaní kompromisného súdu bolo vydané vo forme kráľovského ediktu v roku 1834.

7 Slovenský národný archív v Bratislave, špecializované pracovisko Slovenský banský archív v Banskej Štiavnici, fond Banská komora v Banskej Bystrici, inv. č. 90.

8 Štátny archív v Banskej Bystrici, fond Mesto Banská Bystrica, príručná knižnica, ev. č. 134.


Nos Universitas Prae-
latorum, Baronum, Magnatum et No-
bilium Comitatus Zoliensis damus pro memoria tenore
praesentium significantes, quibus expedit, universis: Quod
cum nos Anno, Die, Mense, Locoque datarum praesentium in-
frascriptis, pro pertractandis altissimum Servitium Regium
ac concomitanter commune hujus Comitatus Bonum tangen-
tibus negotiis in Generali Congregatione convenissemus,
Eotum reliquorum Tractatuum sub decursu, Ioannes [illegible]
Camerae Montanae Neosoliensis, et Regio Cameralis
Dominii Lypcse - Ordinarius Fiscalis nostro exurgens in
Consessu vivae vocis adminiculo super eo apud nos institerit,
quatenus eidem Seriem Metalis Reambulationis Terreni
Liberae Regiae, ac Montanae Civitatis Neosoliensis nobis
sub hodierno relatae in genuino Transumto sub Authenti-
co Comitatus nostri Sigillo expediri facere dignaremur.
Cujus justo petito annuentes praememoratam Reambulati-
onem Metalem Transumi, eidemque una cum Territoriali-
bus Mappis extradari fecimus, in haec verba:
Nos Subscripti mediante gratiosa Inclytae Universitatis
dto 7 Junii 1820 ad Num. Protocolli 598 perlata Determinatio-
ne scopo Reambulationis Metarum, totius Liberae Regiae
ac Montanae Civitatis Neosoliensis Terreni admissi hu-
millime referrimus:

Obrázok 37 Úvodná strana banskoštiavnického exempláru reambulačného protokolu.

Z obsahovej stránky dokument pozostáva z textovej a kartografickej časti. Textová časť obsahuje podrobný opis priebehu obchôdzky hraníc, ktorá prebehla v letných mesiacoch roku 1820, od 1. augusta do 19. septembra. Okrem priebehu obchôdzky je súčasťou textovej časti aj podrobný popis hraníc s jednotlivými hraničnými medzníkmi a znakmi. Celý text je rozdelený na sedem kapitol podľa jednotlivých hraničných úsekov. Prvý popísaný hraničný úsek predstavuje spoločnú hranicu medzi mestom a panstvom rodiny Radvanských. Nasledujú hraničné úseky, na ktorých mesto susedilo s mestom Kremnica, s panstvom rodiny Révaiovcov, s Likavským panstvom, Ľupčianskym panstvom, s panstvom Benických v Mičinej a komposesorátnym územím dediny Iliaš. Hraničné úseky s Kremnicou, panstvom Révaiovcov a Likavským panstvom predstavovali zároveň aj stoličné hranice medzi

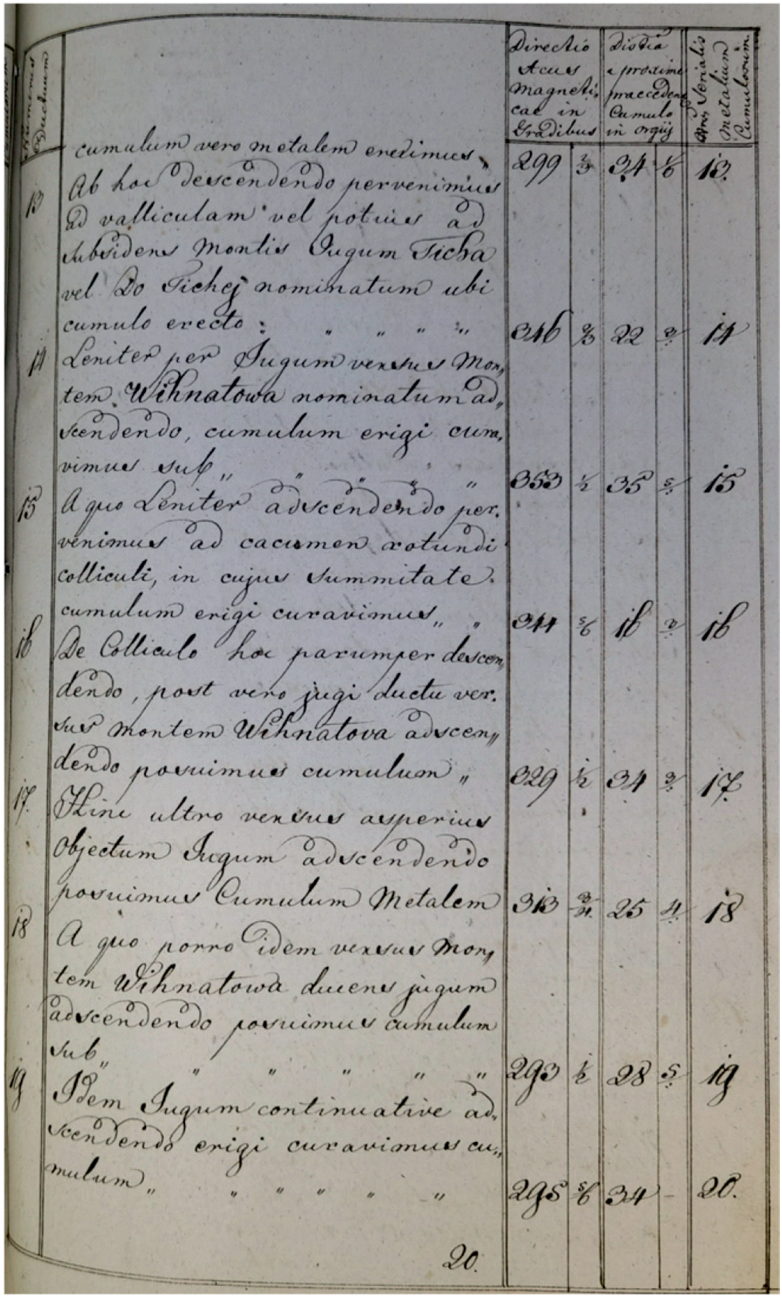
Zvolenskou stolicou a stolicami Tekov, Turiec a Liptov. Súčasťou reambulácie sa nestal jedine popis severovýchodného úseku hraníc mesta v oblasti okolo donovalského sedla, kde mesto hraničilo s Liptovskou stolicou a Likavským hradným panstvom. Tento úsek hranice bol už dlhší čas predtým braný ako sporný,[9] preto bol z reambulácie úplne vynechaný.

Jednotlivé kapitoly majú homogénnu štruktúru. Na začiatku každej kapitoly sa nachádza stručná informácia o popisovanom hraničnom úseku a susediacich subjektoch. Hneď potom nasleduje menný zoznam všetkých účastníkov reambulácie. Pri jednotlivých menách osôb sa uvádzala funkcia, ktorú zastávali v mestskej správe alebo v rámci správy príslušného panstva, respektíve komorskej správy. V prípade mešťanov, ktorí nemali žiadnu funkciu v rámci mestskej správy, sa uvádzal len ich spoločenský status mešťana (*civis*). Osobitne sa uvádzali poddaní a zamestnanci dotknutých subjektov. Medzi účastníkmi reambulácie možno nájsť tiež deti a mládež, teda osoby mladšie ako 18 rokov. V ich prípade sa väčšinou uvádzal aj ich vek a prípadné príbuzenstvo voči dospelým účastníkom hraničnej obchôdzky. Prítomnosť detí a mládeže pri hraničných pochôdzkach a obchádzaní chotárov bolo bežným javom už od čias stredoveku. Ich účasť mala zabezpečiť dlhodobé udržanie vedomostí o priebehu hraníc a umiestnení hraničných medzníkov v historickej pamäti komunity. Na reambulácii každého hraničného úseku sa zrejme zúčastnili aj geometer (zememerač) a osoba zodpovedná za obnovu starých a vytesanie nových hraničných znakov. Účasť geometra počas celej reambulácie možno predpokladať, hoci sa medzi účastníkmi obchôdzky jednotlivých hraničných úsekov vyslovene neuvádza. Vytesanie hraničných znakov mali za úlohu murársky predák Matej Stogl (dva hraničné úseky) a kamenársky majster Leonard Horn (rovnako dva hraničné úseky). V prípade posledných troch hraničných úsekov sa žiadna zmienka o osobe zodpovednej za vytesanie hraničných znakov nenachádza.

Hlavnú a najobsiahlejšiu časť každej kapitoly predstavuje vlastný opis priebehu hranice aj s opisom jednotlivých hraničných medzníkov a znakov. Uvedená časť je spracovaná v tabuľkovej forme s členením na štyri až päť stĺpcov. V prvom stĺpci sa nachádza číselný údaj vyjadrujúci poradie opisovaného hraničného úseku. V najširšom stĺpci sa nachádza písomný opis priebehu hranice a jednotlivých hraničných medzníkov a znakov. V ďalších dvoch stĺpcoch sa nachádzajú číselné údaje, ktoré udávajú uhly a vzdialenosti medzi najbližšími hraničnými medzníkmi. Uhly sa vyjadrovali v stupňoch (*Directio acus Magneticae in Gradibus*) a vzdialenosti vo viedenských siahach (*Distantia in Orgiis Viennensibus*). Vo väčšine kapitol je súčasťou tabuľky aj piaty stĺpec, v ktorom sa uvádzalo číslo odkazujúce na nadväzujúci hraničný úsek. Usporiadanie stĺpcov tabuľky je vo všetkých kapitolách, s výnimkou prvej, zhodné. Po stĺpci s poradím hraničného úseku nasleduje stĺpec s písomným opisom hranice, stĺpec vyjadrujúci uhly medzi medzníkmi, stĺpec vyjadrujúci vzdialenosti medzi medzníkmi a stĺpec s odkazom na číselné poradie nasledujúceho hraničného úseku. Prvá kapitola je ako jediná členená len na štyri stĺpce, pričom na úvod sú zaradené tie, ktoré uvádzajú číselné údaje (poradie úseku, uhly a vzdialenosti medzi medzníkmi). Posledný stĺpec poskytuje priestor na písomný opis príslušného

9 TOMEČEK, Oto: *Drevorubači a uhliari v lesoch Banskej Bystrice*, s. 19 – 21.

hraničného úseku. Na úplnom konci textovej časti kapitoly sa nachádzajú informácie týkajúce sa overenia priebehu remabulácie a reambulovaných hraničných úsekov zo strany stoličných úradníkov.



Obrázok 38 Ukážka tabuľkovej formy zápisu reambulačného protokolu.

Súčasťou dokumentu sú aj mapové prílohy zachytávajúce jednotlivé popísané hraničné úseky. Podobne ako je dokument rozdelený na sedem kapitol zachytávajúcich sedem hraničných úsekov, je súčasťou dokumentu aj sedem máp. Napriek zhodnému počtu kapitol a máp nie všetky úseky zachytené na mape presne zodpovedajú úsekom popísaným v jednotlivých kapitolách. Hranica s Ľupčianskym hradným panstvom je zachytená na dvoch mapách. Naopak dva posledné hraničné úseky s Mičinou a Iliašom sú zachytené len na jednej mape. Kartografické zobrazenie ostatných hraničných úsekov korešponduje s príslušnou kapitolou zachytávajúcou daný hraničný úsek. Jednotlivé mapy sú čiastočne kolorované, predovšetkým hraničná línia, v niektorých prípadoch aj lúky a stavebné objekty stojace v blízkosti hraničnej línie. Orientácia máp nie je jednotná. Pomôckou pri orientácii na každej mape sú vyznačené línie, ktoré označujú smer poludníka (*Linea Meridionalis*), a línia označujúca magnetický sever (*Linea Magnetica*). Pomôckou na meranie vzdialeností na mapách je grafická mierka, ktorá predstavuje vzdialenosť rovnajúcu sa 500 viedenským siaham. Všetky mapy vyhotovil ešte v priebehu roku 1820 riadny stoličný geometer Juraj Mihalko.

Reambulačný protokol pozostávajúci z textovej a mapovej časti predstavuje zásadný dokument umožňujúci presnú rekonštrukciu priebehu historických hraníc mesta Banská Bystrica. Okrem rekonštrukcie samotnej hraničnej línie poskytuje relatívne presnú informáciu o polohe hraničných medzníkov a typológii používaných hraničných znakov. V dokumente sa nachádzajú aj mnohé cenné informácie o historickej krajine lokalizovanej pozdĺž mestských hraníc. Medzi týmito informáciami možno nájsť napríklad údaje o významnejších stavbách ležiacich v blízkosti hraníc. Cenné sú údaje o mostoch a cestách, teda o dobovej cestnej infraštruktúre. Čiastočná pozornosť je venovaná aj polohe, využitiu a majetkovým pomerom pozemkov ležiacich v blízkosti hraničnej línie. Mimoriadne cenné sú početné údaje zaznamenávajúce dobovú toponymiu (hydronymá, oronymá, ojkonymá a iné chotárne názvy). Mnohé z nich sú najstaršími známymi písomnými dokladmi miestneho názvoslovia.

Čistopis predmetného banskoštiavnického exemplára reambulačného protokolu bol napísaný jednou pisárskou rukou v rozmedzí rokov 1820 – 1823. Text rukopisu prečítal a jeho pravosť overil prísažný prvý podnotár (*Iur.[atus] Pr.[imarius] V.[ice] Notarius*) Zvolenskej stolice Ľudovít Radvanský (*Ludovicus de Radvány*). Zápis o tom bol daný na generálnej kongregácii Zvolenskej stolice konanej v Banskej Bystrici dňa 1. októbra 1823.

Opisovaný dokument reambulačného protokolu je zviazaný v polokoženej väzbe s rozmermi 26 x 41 cm. Hrúbka chrbta väzby je približne 5 cm. Celý dokument má rozsah 254 číslovaných strán vrátane zložených mapových listov. Paginácia pôvodne nestránkovaného dokumentu bola vyhotovená až dodatočne, zrejme pri jeho archívnej evidencii a spracovaní. Dodatočne vytvorená paginácia má viacero chýb. Podľa nej má celý dokument len 244 číslovaných strán. Chybná paginácia sa začína na strane 40, ktorá je takto označená dvakrát, naopak strana 83 je vynechaná úplne. Po strane 165 nasleduje strana 156, takže strany 156 až 165 sú chybne zdvojené a strany 156, 157, 158, 159, 160, 161, 162, 163, 164 a 165 sa v dokumente nachádzajú

dvakrát. Podobne dvakrát je očíslovaná aj strana 237. Po odrátaní mapových listov a číslovaných prázdnych strán zostáva textová časť dokumentu, ktorá má 245 strán.

Jazykom celého dokumentu je latinčina. Viaceré názvy toponým sú zapísané v slovenskej podobe, pri niektorých z nich je výnimočne uvedený aj alternatívny nemecký variant názvu. Rukopis celého dokumentu je napísaný pomerne úhľadnou novolatinskou kurzívou. Čitateľnosť rukopisu komplikuje fakt, že pisár použil dve rôzne podoby zápisu malého písmena „s“. Okrem bežnej písanej litery „s“ používal často aj ostré „s“, ktorého čítanie sa zamieňa s podobným zápisom malého písmena „f“. Rôznym spôsobom je zapisované malé písmeno „r“ v prípadoch, keď sa nachádza na začiatku, uprostred alebo na konci slova. Na označenie rozdeleného slova na konci riadka pisár používal pravidelne dve čiarky zapísané na spôsob spodných úvodzoviek. Pomerne ťažko sa v texte rozlišuje písanie malých a veľkých písmen. Skratky sú použité hlavne pri zapisovaní číslovieк, pri ktorých pisár často používal kombináciu arabskej číslice s koncovkou zobranou zo slovného zápisu číslovky. Iné skracovanie slov je skôr výnimkou. Stretnúť sa s ním môžeme pri zápisoch niektorých slov, ako napr. colonus (*col:*), orgia (*org:*), comitatus (*cottus*), civitatis (*cittis*), cameralis (*caalis*), ordinarius (*ord:*), iuratus (*iur:*), manu propria (*m. p.*) alebo Locus Sigilli (*L. S.*). Špecifikom textu je používanie grafických značiek pri zadefinovaní podoby vybraných hraničných znakov. Pri ich tvorbe sa väčšinou používali kapitály písmen. Niektoré hraničné znaky však boli vytvorené aj ako odvodeniny z erbov. Takýto hraničný znak používali napríklad mestá Banská Bystrica a Kremnica. Banská Bystrica väčšinou používala redukovanú podobu svojho erbu[10] predstavujúcu trikrát vodorovne delený štít so štyrmi pruhmi. Túto symboliku mesta v niektorých ojedinelých prípadoch dopĺňali aj majuskulné litery „CN“ (skratka od Civitas Neosolium). Kremnica používala na označenie svojich hraníc katarínske polkoleso, niekedy doplnené kapitálou písmena „C“ (skratka od Cremnitzium).[11] Pisár pri objasňovaní hraničného znaku použitého v teréne zakreslil z dôvodu lepšej názornosti na niektorých miestach takúto erbovú značku aj priamo do samotného rukopisu.

Vzhľadom na pomerne veľký rozsah dokumentu napísaného jednou pisárskou rukou predstavuje dokument vhodnú voľbu na experimentálny prepis metódou automatickej transkripcie. Pozitívom je tiež fakt, že dokument predstavuje čistopis, ktorý obsahuje len minimum škrtov, dodatočných vsuviek a marginálií. Výhodou pre automatickú transkripciu textu je tiež nie príliš široká použitá slovná zásoba a časté opakovanie viacerých slov. Za zásadnejšiu nevýhodu možno na druhej strane považovať tabuľkovú formu spracovania textovej časti dokumentu s nejednotným vzorom tabuľky.

10 Originálny erb Banskej Bystrice mal pôvodne deväť, neskôr osem vodorovných (vo farebnom vyhotovení červeno-bielych) pruhov. Porovnaj: GRAUS, Igor: K najstaršej podobe erbu Banskej Bystrice. In: *Genealogicko-heraldický hlas*, roč. 10, č. 2, 2000, s. 16 – 22; GRAUS, Igor: *Banská Bystrica v 16. storočí: štúdie z dejín mesta*. Banská Bystrica : Enterprise, 2015, s. 152 – 157; NOVÁK, Jozef: *Slovenské mestské a obecné erby*. Bratislava : Slovenská archívna správa, 1967, s. 92 – 95; VRTEĽ, Ladislav: *Osem storočí slovenskej heraldiky*. Martin : Matica slovenská, 1999, s. 63.

11 Pôvodný erb Kremnice známy z 15. storočia obsahoval len zlaté katarínske polkoleso umiestnené v modrom štíte s písmenom „C“ umiestneným nad štítom. Trikrát delený erb obsahujúci katarínske polkoleso s písmenom „C“ v hornom poli, ako aj pruhy a ľalie v dolnej štiepenej polovici erbu vznikol v 16. storočí. Porovnaj: NOVÁK, J.: *Slovenské mestské a obecné erby*, s. 126 – 128; VRTEĽ, L.: *Osem storočí slovenskej heraldiky*, s. 137 – 138.

PRÁCA S DOKUMENTOM V PLATFORME *TRANSKRIBUS*

Transkribus je jedným z viacerých dostupných nástrojov, ktoré je možné využiť na automatickú transkripciu historických rukopisných textov. Medzi takéto aplikácie a nástroje určené na automatickú transkripciu patria napríklad: eScriptorium, OCR4all, PERO – OCR a iné. Platformu *Transkribus* vyvinulo konzorcium pod vedením Güntera Mühlbergera z Univerzity v Innsbrucku v rámci riešenia projektu Horizont 2020 READ (Recognition and Enrichment of Archival Documents). V porovnaní s ostatnými platformami určenými na automatickú transkripciu rukopisných textov je jedinou, ktorá umožňuje bežným používateľom vytvárať a trénovať vlastné modely schopné rozpoznávať a prepisovať rukopisy.[12] Platforma sa medzičasom skomercionalizovala, avšak naďalej sa technologicky vyvíja a zdokonaľuje. Možnosť vytvoriť si vlastný model ušitý na mieru konkrétnemu dokumentu, rovnako ako aj možnosť jeho ďalšieho vylepšovania dodatočným trénovaním zvyšujú atraktivitu nástroja a je hlavným dôvodom jeho uprednostnenia riešiteľmi projektu APVV SKRIPTOR pred ostatnými podobnými aplikáciami a nástrojmi.

Prvým nevyhnutným krokom predchádzajúcim samotnej práci v systéme *Transkribus* je primárna digitalizácia vybraného dokumentu. Ideálnym spôsobom zosnímania a transformácie dokumentu do digitálnej podoby je jeho naskenovanie prostredníctvom výkonného skenera. V prípade reambulačného protokolu, ktorý doposiaľ nebol zdigitalizovaný, som zvolil jeho nafotenie prostredníctvom bežného fotoaparátu, ktorý je súčasťou mobilného telefónu (smartphonu). Projekt READ a ním vytvorená platforma *Transkribus* počíta s možnosťou núdzového nasnímania archívnych dokumentov prostredníctvom smartphonov. Práve na tento účel bola vyvinutá pomôcka ScanTent a softvér DocScan app pre Android. Prenosný ScanTent disponuje vlastným podsvietením, ktoré zabezpečuje rovnomerné osvetlenie dokumentu a eliminuje jeho možné zatienenie počas jeho nasnímania. Obsahuje tiež pevnú podložku na umiestnenie smartphonu, ktorá je kompatibilná so všetkými typmi smartphonov. Nevýhodou ScanTentu je obmedzený rozmer plochy určenej na umiestnenie snímaného dokumentu. Jej rozmery umožňujú pohodlné zosnímanie dokumentov len do veľkosti A3, respektíve len mierne presahujúcej formát A3. Softvér DocScan app, ktorý je možné voľne stiahnuť do mobilného smartphonu, umožňuje zosnímanie dokumentu aj bez neustáleho stláčania spúšte fotoaparátu. Bádateľ sa tak môže sústrediť na prácu s dokumentom a pretáčanie jeho strán alebo prípadnú výmenu voľných listov.[13]

Nasnímanie celého rukopisu banskoštiavnického exempláru reambulačného protokolu prebehlo v študovni Slovenského banského archívu v Banskej Štiavnici dňa 25. septembra 2020. Pri nasnímaní dokumentu som využil ScanTent a dva mobilné telefóny (Huawei P40 Pro a Google Pixel 4), oba bez aplikácie DocScan app. Napriek nutnosti manuálneho spúšťania fotoaparátov mobilných telefónov

12 KATUŠČÁK, Dušan: Digital humanities a automatická transkripcia rukopisných textov. In: *ITlib: informačné technológie a knižnice*, roč. 24, č. 1, 2020, s. 9.

13 K obom uvedeným pomôckam pozri viac na webstránke: The ScanTent: Professional scanning with your smartphone. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-10-28]. Dostupné na: https://readcoop.eu/scantent/

trvalo nasnímanie celého dokumentu (129 dvojstrán) na jeden prístroj približne 25 minút. Dve nasnímania celého dokumentu na dva rôzne mobilné telefóny teda trvali necelú jednu hodinu.

Po nasnímaní dokumentu som obe verzie fotografií stiahol do osobného počítača s operačným systémom Windows. Snímky vo formáte JPEG som následne roztriedil podľa ich kvality. Tu sa ukázala výhoda nasnímania dokumentu na dva telefóny a vytvorenie záložného súboru snímok. Keďže niektoré snímky nedosahovali dostatočnú kvalitu, mohli byť nahradené snímkami vytvorenými druhým telefónom zo záložného súboru. Za základ finálneho súboru snímok som zvolil zábery vyhotovené mobilným telefónom Huawei P40 Pro. Tieto boli vyhotovené s rozlíšením 96 dpi a rozmermi 4096 x 3072 pixelov. V prípade nutnosti som tieto snímky nahradil zábermi vytvorenými mobilným telefónom Google Pixel 4, ktoré boli vyhotovené s nižším rozlíšením 72 dpi a s rozmermi 4032 x 3024 pixelov. Nahradenie snímok bolo nevyhnutné len v 13 prípadoch. Išlo o snímky, ktoré neboli dostatočne ostré, prípadne obsahovali viaceré rozmazané znaky. Finálny súbor pripravený na importovanie do systému *Transkribus* tak tvorilo 116 snímok dvojstrán vyhotovených telefónom Huawei v rozlíšení 96 dpi a 13 snímok dvojstrán vyhotovených telefónom Google Pixel v rozlíšení 72 dpi.

Po primárnej digitalizácii a roztriedení jednotlivých záberov podľa kvality bolo možné pristúpiť k importovaniu pripravených snímok do systému *Transkribus*. Samozrejme ešte predtým bolo potrebné vytvoriť si vlastné konto po zaregistrovaní na stránke spoločnosti READ-COOP SCE.[14] Následne po stiahnutí platformy do osobného počítača a vytvorení vlastného účtu[15] mohlo prísť k založeniu vlastnej zbierky (collection) pod názvom Metales_SNASBA a zložky (folder) určenej na umiestnenie digitalizátov. Celá zložka so súborom všetkých digitalizátov, obsahujúca 129 obrázkov vo formáte JPEG predstavujúcich 129 dvojstrán a 258 samostatných strán dokumentu (z toho text sa nachádza na 245 stranách), mala veľkosť 501 MB. Zložku s obrázkami tejto veľkosti bolo možné importovať do systému jednorazovo bez potreby rozdeľovať súbor na viacero menších súborov.

Po importovaní súborov sa všetky ďalšie práce spojené s automatickou transkripciou rukopisného textu realizujú už len v online prostredí platformy *Transkribus*. V prvej fáze je potrebné textové polia na každej snímke segmentovať, teda vyznačiť tie bloky textu a jednotlivé riadky, ktoré majú byť predmetom automatickej transkripcie. Súčasťou segmentácie je aj určenie poradia čítania jednotlivých textových blokov a riadkov. Segmentáciu je možné vykonať automaticky alebo manuálne. Automatická segmentácia je vhodná najmä pre jednoduchý, vnútorne nečlenený text. Po automatickej segmentácii je potrebné vykonať kontrolu segmentovaného dokumentu a opraviť prípadné chyby. Nie príliš vhodná je automatická segmentácia pre dokumenty vytvorené v tabuľkovej podobe, medzi ktoré možno zaradiť aj zvolený reambulačný protokol.

14 Spoločnosť READ-COOP SCE, s. r. o., bola založená 1. júla 2019 s cieľom udržať a ďalej rozvíjať platformu *Transkribus* vyvinutú v rámci riešenia projektu Horizont 2020 READ.

15 Zriadenie účtu, prihlásenie a stiahnutie platformy Transkribus je možné na webovej stránke spoločnosti READ-COOP SCE. Pozri: Download *Transkribus*. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-10-28]. Dostupné na: https://readcoop.eu/transkribus/download/

Napriek uvedenému som sa rozhodol celý dokument nechať najprv prejsť automatickou segmentáciou. Na tých stranách dokumentu, ktoré neobsahovali tabuľky, ale len samotný text, prebehla segmentácia bez väčších problémov. Takýchto strán však obsahuje dokument len minimum, keďže prevažná časť dokumentu je spracovaná v tabuľkovej podobe. V prípade strán vytvorených tabuľkovou formou sa už automatická segmentácia ukázala ako nie príliš vhodný nástroj úpravy dokumentu. Po automatickej segmentácii tieto strany obsahovali príliš veľa chýb. Niektoré znaky boli vyhodnotené nepresne. Ako samostatné znaky boli napríklad označené aj také časti dokumentu, ktoré v skutočnosti neboli alfanumerickými znakmi. Išlo o rôzne škvrny či znaky presvitajúce z opačnej strany. Naopak, niektoré existujúce znaky, ktoré mal *Transkribus* označiť, zostali neidentifikované a figurovali v dokumente ako neexistujúce alfanumerické znaky. Pri stranách vytvorených v tabuľkovej podobe si *Transkribus* väčšinou neporadil ani s určením poradia čítania jednotlivých blokov a riadkov.

Všetky uvedené chyby, ktoré vznikli počas automatickej segmentácie, bolo potrebné odstrániť manuálnou segmentáciou jednotlivých strán. Pri manuálnej segmentácii som realizoval segmentáciu so zachovaním tabuľky, ako aj segmentáciu bez tabuľky. V prípade strán, pri ktorých som sa rozhodol zachovať tabuľkovú podobu, som najprv vymazal pôvodnú automatickú segmentáciu danej strany. Následne prostredníctvom horizontálneho a vertikálneho delenia strany som vyznačil jednotlivé riadky a stĺpce. Takto došlo k rozdeleniu tabuľky na niekoľko buniek. Vo forme tabuľky upravená strana sa mohla opätovne segmentovať. Po jej segmentovaní bolo potrebné skontrolovať a v prípade potreby upraviť poradie čítania jednotlivých buniek tabuľky, prípadne poradie čítania jednotlivých riadkov v rámci týchto buniek.

Tento postup sa ukázal pomerne zdĺhavý. Z uvedeného dôvodu som sa rozhodol pri ďalšej úprave dokumentu upustiť od zachovania tabuliek pri segmentácii. Situáciu som vyhodnotil tak, že jediný potrebný údaj na správne dešifrovanie obsahu dokumentu a jeho ďalšie praktické využitie v rámci historického výskumu je okrem textového poľa len stĺpec, ktorý označuje poradie opisovaného hraničného úseku. Stĺpce s údajmi o uhle a vzdialenosti medzi jednotlivými hraničnými medzníkmi, ako aj o poradí nadväzujúceho hraničného úseku nie sú na pochopenie dokumentu až také podstatné. Z tohto dôvodu som sa napokon rozhodol tieto údaje do automatickej transkripcie nezaradiť. K uvedenému rozhodnutiu prispela aj vedomosť, že podstatná časť z predmetných medzníkov sa doposiaľ nezachovala, a preto by bolo pomerne náročné rekonštruovať presný priebeh hraničnej línie len na základe údajov o vzdialenostiach a uhloch. Druhým dôvodom tohto rozhodnutia bola praktická skúsenosť s tým, že *Transkribus* má podstatne väčší problém s rozpoznaním a správnym prepisom numerických znakov v porovnaní s písmenami.

Vzhľadom na tieto fakty nebolo potrebné vytvárať pri segmentovaní dokumentu tabuľku. Úprava textu vnímaného ako jeden celok potom vyzerala tak, že číselný údaj o poradí konkrétneho hraničného úseku som označil ako samostatný riadok, pod ktorým pokračoval text popisujúci hraničný úsek označený daným číslom. Po tomto rozhodnutí som automaticky segmentované strany už nezmazával, aby som

ich po nevyhnutnej úprave opätovne segmentoval, ale len upravoval do vyššie načrtnutej podoby. Úprava takejto automaticky segmentovanej strany pozostávala z niekoľkých nevyhnutných krokov. Najprv som odstránil všetky chybne označené znaky a bloky textu. Nasledovala kontrola a prípadná korekcia poradia čítania jednotlivých riadkov. Posledným krokom bola kontrola spresnenia hranice dĺžky každého riadku.

Transkribus pri segmentovaní využíva spravidla tri prednastavené farby – zelenú, fialovú a modrú. Zelená farba označuje jednotlivé bloky textu. V prípade vnútorne nečleneného textu sa tento blok väčšinou prekrýva s rozsahom celej strany. V prípade tabuľkovej formy označujú tieto bloky jednotlivé polia tabuľky. Fialová farba označuje dĺžku riadku od jeho začiatku do konca. Modrá farba slúži na označenie jeho výšky. Vzhľadom na fakt, že rukopis obsahuje malé aj veľké písmená, pričom niektoré písmená presahujú výšku riadku (v niektorých prípadoch dokonca prechádzajú do susedného riadku), presné vyznačenie výšky riadku si vyžaduje dodatočnú úpravu. Tú je možné urobiť vytváraním polygónov okolo jednotlivých písmen a slov v riadku. Takýmto spôsobom možno veľmi presne ohraničiť dĺžku aj výšku riadku s presným zadefinovaním všetkých znakov prináležiacich do konkrétneho riadku. Vytváranie takýchto polygónov, ohraničujúcich textové pole, sa ukázalo byť časovo mimoriadne náročné. Úprava jednej dvojstrany pozostávajúcej z cca 60 – 70 riadkov (na jednej strane sa zvačajne nachádza text zapísaný v 30 až 35 riadkoch) trvala asi jednu celú hodinu. Po úprave prvých 17 dvojstrán, predstavujúcich 32 jednotlivých strán dokumentu, som od tohoto spôsobu úpravy dokumentu nakoniec upustil. Pri ďalšej úprave nasledovných segmentovaných strán som sústredil svoju pozornosť len na odstránenie vyznačených nepotrebných znakov, vyznačenie blokov textového poľa, úpravu poradia čítania jednotlivých riadkov a kontrolu správnej dĺžky riadkov. Celý proces úpravy jednej dvojstrany sa takto skrátil na približne 20 – 25 minút, čo v porovnaní s predošlým spôsobom úpravy jednej dvojstrany predstavovalo časovú úsporu 35 – 40 minút.

Dĺžka segmentácie môže závisieť od rozsahu, úpravy a celkového charakteru textového poľa, počtu riadkov na jednotlivých stranách dokumentu, osobitných zručností každého jednotlivca, ako aj od aktuálnej fázy segmentácie dokumentu. V počiatkoch segmentácie môže byť postup pomalší, avšak po jeho zautomatizovaní sa tento proces môže postupne zrýchľovať. I. Nagy napríklad uvádza časovú náročnosť na segmentovanie jednej dvojstrany len na úrovni približne desiatich minút.[16]

Po ukončení segmentácie sa mohla začať ďalšia etapa práce, ktorú predstavuje manuálny prepis vybranej časti dokumentu. Tento prepis možno zrealizovať priamo do platformy *Transkribus* alebo do osobitného súboru, z ktorého je potom potrebné prepísaný text nakopírovať do systému. Ja som zvolil prepis na nečisto do súboru vo formáte Word. Hlavným dôvodom tohto rozhodnutia bolo ponechať si možnosť

16 NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 58; NAGY, Imrich: Transkribus ako nástroj na sprístupnenie dobových archívnych pomôcok na príklade Csákósovho katalógu korešpondencie Kohárovcov. In: *Digital humanities: nástroje sprístupňovania historického dedičstva. Zborník abstraktov*. eds. P. Maliniak – I. Nagy, Banská Bystrica : Štátna vedecká knižnica, 2022, s. 66.

realizovať prípadné dodatočné korekcie prepísaného textu aj pomocou fulltextového vyhľadávania jednotlivých slov. Takto prepísaný dokument je možné neskôr veľmi jednoducho skopírovať a preniesť do systému *Transkribus*.

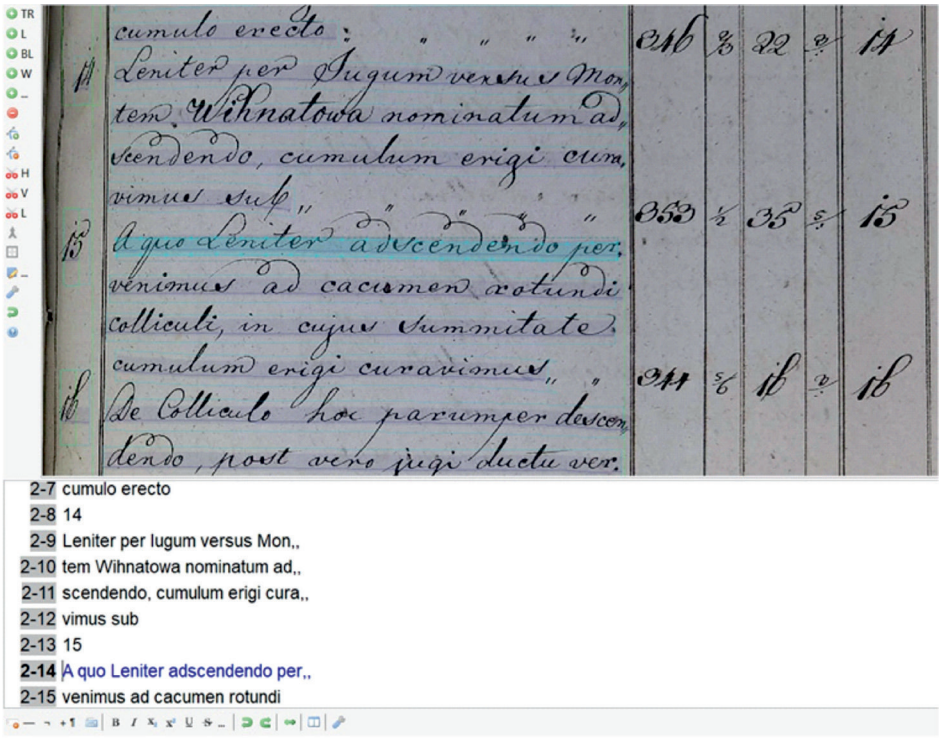
Na manuálny prepis je potrebné vybrať určitú vzorku dokumentu, na ktorej je potom potrebné trénovaním vytvoriť vlastný model. Takýto model sa neskôr použije na finálny automatický prepis zostávajúcej časti dokumentu, ktorá dovtedy nebola manuálne prepísaná. Spomedzi 245 textových strán celého dokumentu som sa rozhodol zrealizovať manuálny prepis prvých 49 strán textu. Uvedená voľba predstavovala približne jednu pätinu z celkového rozsahu textovej časti dokumentu. Táto vzorka sa neskôr ukázala ako nedostatočná, pretože ju tvorilo len necelých 7 000 slov. Autori *Transkribu* pritom odporúčajú pripraviť na vytrénovanie vlastného modelu vzorku textu, ktorý obsahuje aspoň okolo 15 000 slov.[17] Napriek uvedenému, som sa rozhodol vyskúšať možnosti automatickej transkripcie aj na príklade uvedenej vzorky dokumentu s menším rozsahom prepísaných slov.

Pri manuálnom prepise vybranej časti dokumentu bolo potrebné postupovať metódou transliterácie. Vo vlastnom prepise teda nebolo možné opravovať žiadne chyby ani rozpisovať žiadne textové skratky. Ako príklad pisárskej chyby možno uviesť nesprávny zápis slova *nominalis*, namiesto správneho *nominatis*.[18] Veľmi nedôsledne pisár dokumentu zapisoval malé a veľké písmená. Aj v tomto prípade bolo potrebné zachovať pôvodný zápis, hoci v tejto súvislosti je potrebné uviesť, že rozlišovanie malých a veľkých písmen bolo na viacerých miestach značne problematické. Jedinou výnimkou, pri ktorej som nedodržal zásadu presnej transliterácie, bolo vynechanie zapisovania ostrého „s“. Hoci pisár zapisoval písmeno „s“ dvojakým spôsobom, použil som pri prepise len jednoduchý zápis tohto písmena. Ako príklad možno uviesť slovo *proceßßimus*, ktoré bolo v mojom prepise zapísané ako *processimus*. V tomto prípade som usúdil, že s jednotným prepisom dvoch rôznych foriem zápisu písmena „s“, by *Transkribus* nemal mať zásadnejší problém. Program by sa mal v tomto prípade jednoducho naučiť priradiť dvom rozdielnym znakom pri prepise len jeden znak. Súčasťou dokumentu je aj niekoľko škrtov, pri ktorých je prečiarknuté celé slovo alebo jeho časť. Keďže *Transkribus* umožňuje vyznačenie preškrtnutých slov, bolo možné v tomto prípade dodržať označenie škrtov aj pri prepise. Potrebné bolo napokon striktne dodržať tiež riadkovanie dokumentu, aby každý prepísaný riadok zodpovedal príslušnému riadku v dokumente.

Po nakopírovaní prepísaných strán do systému *Transkribus* bolo potrebné zrealizovať záverečnú kontrolu, aby bolo jasné, že pod príslušnou stranou sa nachádza príslušný prepis uvedenej strany. Kontrola sa týkala aj overenia vzájomnej komunikácie medzi označenými riadkami originálneho dokumentu so zodpovedajúcimi riadkami prepísaného textu. Všetky takto skontrolované strany som označil ako Ground Truth vzorku, teda finálnu verziu úpravy dvojstrany. Takto označené strany boli pripravené na tvorbu a trénovanie modelu, prípadne samotnú automatickú transkripciu.

17 NAGY, Imrich: Možnosti aplikácie metódy..., s. 56.

18 Takto na konci strany 11 a na začiatku strany 12 v spojení „*... in Collibus holie Huorky nominalis siti...*“



Obrázok 39 Ukážka manuálneho prepisu rukopisného textu v platforme *Transkribus*.

TVORBA MODELOV AUTOMATICKEJ TRANSKRIPCIE

Vlastnej automatickej transkripcii historického rukopisného textu predchádza vytvorenie vlastného modelu, ktorý umožní transkripciu zrealizovať. Pri vytváraní modelu automatickej transkripcie je potrebné manuálne prepísaný text dokumentu rozdeliť na dve časti. Rozsahom väčšia časť predstavuje cvičný súbor (Training Set), zvyšná časť predstavuje overovací súbor (Validation Set). Odporúčaný pomer rozdelenia prepísanej časti dokumentu medzi cvičný a overovací súbor je 10 : 1.[19] Na rozsahom väčšej vzorke textu zaradenej do cvičného súboru sa *Transkribus* učí priraďovať konkrétny alfanumerický znak znaku použitému pri rukopisnom zápise textu. To, čo sa naučí na vzorke cvičného súboru, potom využije pri prepise textu zaradeného do overovacieho súboru. Takýmto spôsobom sa vytvorí model automatickej transkripcie. Smerodajným pri jeho praktickom využití je výsledok dosiahnutý pri prepise v rámci overovacieho súboru.

Transkribus dokáže vyhodnotiť chybovosť na úrovni jednotlivých slov (WER, t. j. Word Error Rate) aj znakov (CER, t. j. Character Error Rate), a to tak v rámci cvičného, ako aj overovacieho súboru. Keďže v prípade cvičného súboru porovnáva *Transkribus* rukopisné

19 NAGY, Imrich: Možnosti aplikácie metódy..., s. 58.

znaky priamo s manuálne prepísanými znakmi, chybovosť dosahuje zákonite nižšie hodnoty ako v prípade overovacieho súboru. Dôležitejším, z hľadiska neskoršieho výberu modelu určeného na finálnu automatickú transkripciu, je však výsledok prepisu textu zaradeného do druhého overovacieho súboru. Podľa údajov D. Katuščáka by sa chybovosť na úrovni slov mala pohybovať pod 30 % a chybovosť na úrovni znakov pod 15 %. Pri dosiahnutí týchto hodnôt je transkribovaný text pre človeka ešte pochopiteľný a použiteľný.[20] I. Nagy hovorí o zmysluplnosti automatickej transkripcie pri dosiahnutí chybovosti na úrovni znakov nižšej ako 10 %. Za vynikajúci sa považuje výsledok nižší ako 5 % na úrovni chybovosti znakov, pričom chybovosť dosahujúcu 1 – 2 % na úrovni znakov je možné dosiahnuť väčšinou len pri tlačených dokumentoch.[21] Vzhľadom na uvedené hodnoty som si stanovil cieľ vytrénovať model pre potreby automatickej transkripcie rukopisného textu reambulačného protokolu, ktorý bude dosahovať chybovosť na úrovni znakov v overovacom súbore v rozmedzí od 3 do 5 %. Vytrénovanie modelu na tejto úrovni znamená de facto úspešnosť prepisu na úrovni znakov dosahujúcu hodnoty 95 – 97 %. Vytrénovanie takéhoto modelu možno považovať za úspešný výsledok a signál na ukončenie trénovania modelu, prípadne za začiatok automatickej transkripcie zvyšnej, manuálne neprepísanej časti rukopisného dokumentu.

Pomôckou pri vylepšovaní (trénovaní) vlastného modelu môže byť základný model (Base Model) dostupný v portfóliu voľne prístupných modelov na platforme *Transkribus*. V tomto prípade ide o model, ktorý na podklade iného rukopisu vytrénoval jeden z používateľov platformy. Využitie základného modelu pri vylepšení vlastného modelu je možné, samozrejme, len v prípade, že je voľne dostupná aj ukážka originálneho rukopisu, na základe ktorého tento základný model vznikol. Vzájomným porovnaním rukopisu základného modelu s rukopisom môjho dokumentu potom možno vybrať najvhodnejší základný model využiteľný na trénovanie vlastného modelu. Táto voľba znamená, že informácie obsiahnuté v základnom modeli budú integrované do nového modelu. Týmto spôsobom je možné významne urýchliť celý proces trénovania vlastného modelu. Využitie základného modelu sa stáva veľkým benefitom najmä v prípade, že máme k dispozícii rozsahom menšiu vzorku textu zaradeného do cvičného súboru.[22] Využitie základného modelu takto dokáže ušetriť čas nevyhnutný na zdĺhavý manuálny prepis a prípravu rozsiahlejšej vzorky.

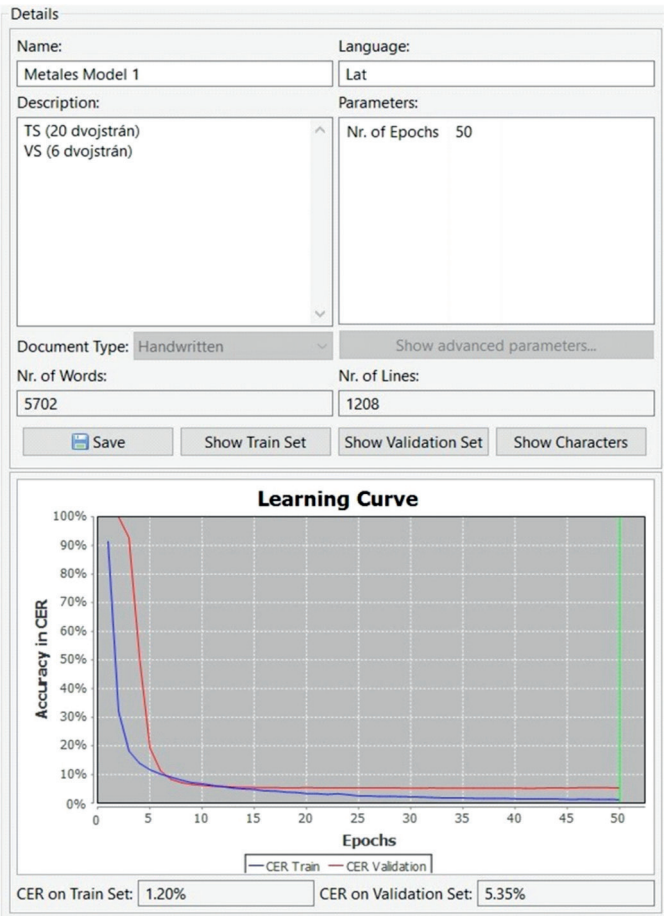
Pre potreby tvorby modelu číslo 1 som mal k dispozícii manuálne prepísaný text zo snímok z prvých 26 dvojstrán dokumentu, čo predstavovalo 49 strán textu so 7 202 slovami. Do cvičného súboru som zaradil text obsahujúci 5 702 slov, pre overovací súbor zostal text obsahujúci 1 500 slov. Namiesto odporúčaného pomeru 10 : 1 medzi cvičným a overovacím súborom som teda svoju vzorku rozdelil približne v pomere 4 : 1. Napriek uvedenému bol výsledok trénovania modelu číslo 1 povzbudivý. V prípade cvičného súboru, na ktorom sa *Transkribus* učí identifikovať jednotlivé znaky, sa podarilo dosiahnuť chybovosť na úrovni znakov 1,20 %. V prípade dôležitejšieho overovacieho súboru, ktorý ukazuje, ako dokáže program prečítať text podľa toho, čo sa naučil na podklade cvičného súboru, sa podarilo dosiahnuť chybovosť na úrovni znakov 5,35 %.

20 KATUŠČÁK, Dušan: Digital humanities..., s. 14.

21 NAGY, Imrich: Možnosti aplikácie metódy..., s. 59 – 60.

22 How To Train and Apply Handwritten Text Recognition Models in Transkribus eXpert. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-10-28]. Dostupné na: https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/

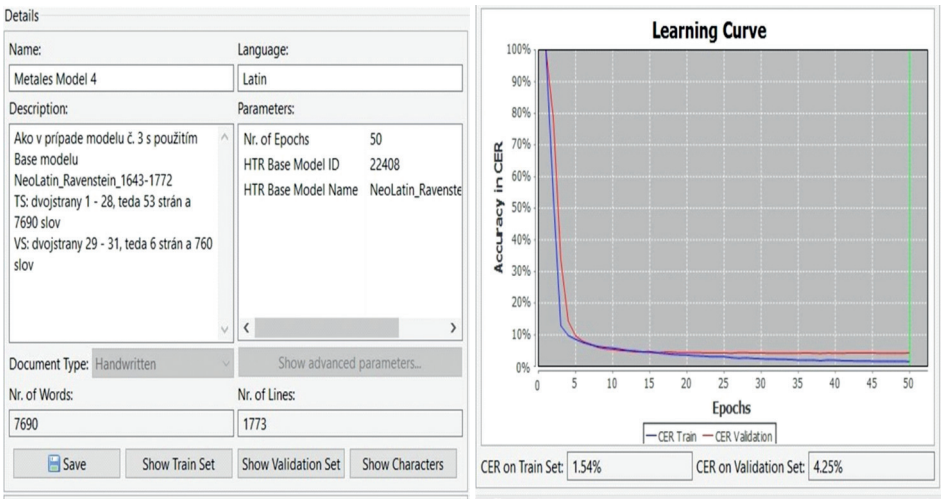
Rovnakú vzorku textu s rovnakým rozdelením medzi cvičný a overovací súbor som sa následne pokúsil vylepšiť využitím základného modelu, teda modelu voľne prístupného na platforme *Transkribus*. Spomedzi dostupných základných modelov sa najviac rukopisu môjho dokumentu približoval taký, na ktorom bol vytrénovaný model s označením Neolatin Ravenstein 1643 – 1772. Tento základný model bol vytrénovaný s chybovosťou na úrovni znakov dosahujúcou hodnotu 4,51 % v cvičnom súbore a 3,58 % v overovacom súbore. Po vylepšení modelu, využitím tohto základného modelu, vznikol vlastný model číslo 2. V prípade tohto nového modelu sa chybovosť na úrovni znakov v cvičnom súbore nepatrne zvýšila na 1,33 %, avšak v dôležitejšom overovacom súbore sa znížila na hodnotu 4,74 %. Výsledok s chybovosťou na úrovni znakov pod 5 % sa pri rukopisoch považuje za veľmi dobrý výsledok. Aj pri vzorke textu s výrazne menším počtom slov v porovnaní s odporúčaným počtom slov, rovnako ako aj s polovičným pomerom medzi cvičným a overovacím súborom v porovnaní s odporúčaným pomerom, sa podarilo dosiahnuť veľmi dobré výsledky. Model číslo 2 teda možno hodnotiť ako plne funkčný a pripravený na automatickú transkripciu rukopisu reambulačného protokolu.



Obrázok 40 Vyhodnotenie trénovania a parametre modelu číslo 1.

Napriek úspešnému vytrénovaniu predošlého modelu som sa rozhodol overiť možnosť jeho ďalšieho vylepšenia dodatočným rozšírením cvičného súboru. Rozšíriť cvičný súbor si vyžadovalo manuálne prepísať ďalšie dovtedy neprepísané strany rukopisu. Dodatočný manuálny prepis zahŕňal päť dvojstrán, na ktorých sa nachádza desať klasických strán textu. Rozšírenie manuálne prepísaného textu umožnilo rozšíriť cvičný súbor na tvorbu modelu číslo 3 na 7 690 slov. Tento počet slov predstavuje približne polovicu z odporúčaného počtu slov na vytrénovanie modelu. Do druhého overovacieho súboru som zaradil zvyšných 760 slov manuálne prepísaného textu. Pri tomto rozdelení slov do jednotlivých súborov som dosiahol odporúčaný pomer 10 : 1. Výsledkom trénovania modelu číslo 3 bola chybovosť na úrovni znakov 1,71 % v cvičnom súbore a 5,09 % v overovacom súbore. Chybovosť v cvičnom súbore sa v porovnaní s predošlými modelmi číslo 1 a 2 paradoxne zvýšila. Chybovosť v overovacom súbore sa v porovnaní s modelom číslo 1 nepatrne znížila, avšak v porovnaní s modelom číslo 2 bola vyššia.

Uvedené výsledky som sa opätovne pokúsil vylepšiť pomocou využitia dostupného základného modelu Neolatin Ravenstein 1643 – 1772. Pri trénovaní nového modelu číslo 4 som vzhľadom na možnosť lepšieho vyhodnotenia experimentu ponechal totožné rozdelenie manuálne prepísaného textu medzi cvičným a overovacím súborom, ako to bolo pri tvorbe predošlého modelu číslo 3. Aj v tomto prípade bolo výsledkom ďalšie vylepšenie modelu. Nový model číslo 4 dosiahol chybovosť na úrovni znakov 1,54 % v cvičnom súbore a 4,25 % v overovacom súbore. Chybovosť v cvičnom súbore sa v porovnaní s modelom číslo 3 znížila, avšak v porovnaní s modelmi 1 a 2 bola o niečo vyššia. Chybovosť v dôležitejšom overovacom súbore dosiahla najnižšie hodnoty v porovnaní so všetkými dovtedy vytvorenými modelmi. Výsledok dosiahnutý v overovacom súbore indikuje, že model číslo 4 možno považovať za zatiaľ najvhodnejší spomedzi všetkých doposiaľ vytvorených modelov, určených na automatickú transkripciu vybraného dokumentu.



Obrázok 41 Vyhodnotenie trénovania a parametre modelu číslo 4.

Tabuľka 2 Trénovanie modelov s porovnaním ich chybovosti na úrovni znakov (CER).

Model	Training Set		Validation Set	
	Nr. of Words	CER	Nr. of Words	CER
Model 1	5 702	1,20 %	1 500	5,35 %
Model 2 (+ Base M.)	5 702	1,33 %	1 500	4,74 %
Model 3	7 690	1,71 %	760	5,09 %
Model 4 (+ Base M.)	7 690	1,54 %	760	4,25 %

Výrazné znižovanie chybovosti na úrovni znakov pri modeloch kombinovaných so základným modelom Neolatin Ravenstein 1643 – 1772 ma priviedlo k myšlienke pokúsiť sa uskutočniť automatický prepis vybranej vzorky vlastného dokumentu výlučne pomocou tohto základného modelu. Na tento experiment som vybral dve vzorky textu s počtom 1 500 slov a 760 slov. Išlo o rovnaké vzorky textu, ktoré som zaradil do overovacieho súboru pri modeloch číslo 1 a 2, respektíve v druhom prípade pri modeloch číslo 3 a 4. V prípade prvej vzorky obsahujúcej 1 500 slov základný model dokázal prepísať text s chybovosťou 72,45 % na úrovni slov a 24,32 % na úrovni znakov. V prípade druhej vzorky obsahujúcej len 760 slov základný model dokázal prepísať text s chybovosťou 79,71 % na úrovni slov a 27,80 % na úrovni znakov. V oboch prípadoch bol výsledkom automatického prepisu text s veľmi vysokou chybovosťou, ktorá indikuje jeho nízku zrozumiteľnosť a obmedzenú použiteľnosť. Záverom tohto experimentu je poznanie, že využitie základného modelu bez použitia vlastného modelu nie je vhodnou metódou automatickej transkripcie vlastného dokumentu. Naopak, ako opodstatnené sa ukázalo vylepšovanie vlastných modelov prostredníctvom základných modelov.

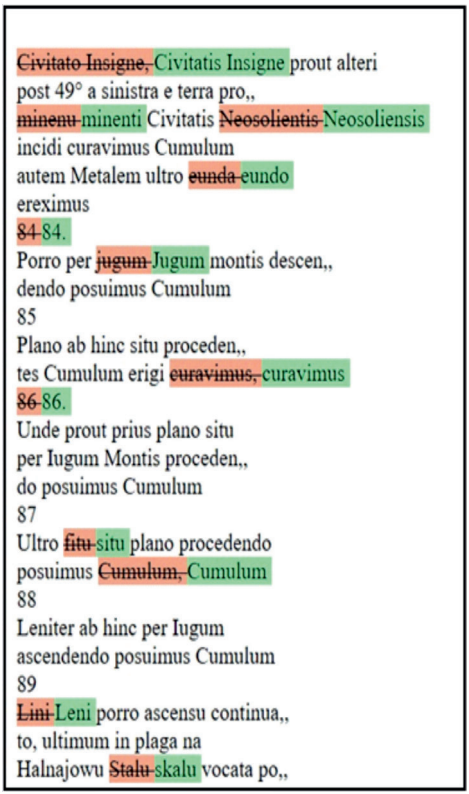
Tabuľka 3 Chybovosť na úrovni slov (WER) a znakov (CER) dosiahnutá pri prepise časti rukopisu reambulačného protokolu výlučne s použitím základného modelu NeoLatin Ravenstein 1643 – 1772.

Vzorka	Nr. of Words	WER	CER
ako model 1 a 2	1 500	72,45 %	24,32 %
ako model 3 a 4	760	79,71 %	27,80 %

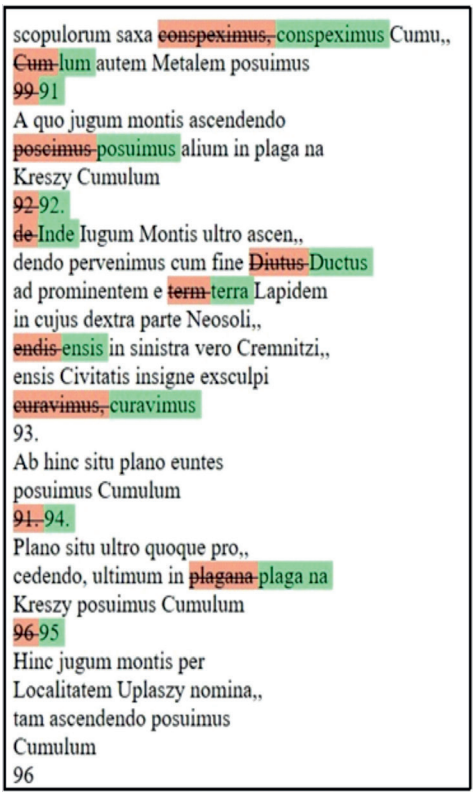
Na príklade modelu číslo 4, dosahujúceho najnižšiu chybovosť na úrovni znakov spomedzi všetkých štyroch vlastných modelov, je možné poukázať na konkrétne chyby vzniknuté pri automatickej transkripcii. Medzi časté chyby možno zaradiť nesprávne identifikovanie veľkých a malých písmen, interpunkcie, číslic alebo medzier medzi slovami. V niektorých prípadoch došlo k nesprávnej identifikácii a následnej zámene podobných písmen. Ako príklad možno uviesť občasnú zámenu dvojíc písmen: *a* –

o, f – s, i – e, c – e, t – k, t – l, t – b, h – l, n – r alebo *r – m*. Asi najčastejšia bola zámena písmen *n* a *r*. Zámena iných písmen bola skôr zriedkavá. Občasné problémy mal *Transkribus* aj s identifikáciou písmena *z*. Ako príklad možno uviesť toponymum *Uplazy*, ktoré v tejto forme zápisu nedokázal *Transkribus* ani raz správne prečítať. Na jednom mieste toto slovo prepísal chybne ako „*Uplamus*“, inde „*Uplans*“ či „*Uplatu*“. Chyby vzniknuté pri rozlišovaní malých a veľkých písmen, ako aj pri písaní interpunkcie možno považovať za menej závažné chyby, ktoré výraznejšie neovplyvnili zrozumiteľnosť transkribovaného textu. Za vážnejšie naopak možno považovať chyby vzniknuté zámenou písmen a nesprávnou identifikáciou arabských číslic. Prekvapivá je najmä mimoriadne vysoká nepresnosť pri identifikovaní číselných znakov. Na šiestich stranách textu zaradeného do overovaciu súboru modelu číslo 4 sa nachádza spolu 55 číselných znakov, z ktorých *Transkribus* dokázal správne identifikovať len 30. Tieto hodnoty prezrádzajú veľmi nízku, približne 55 % úspešnosť pri identifikácii číslic. Napriek všetkým uvedeným chybám možno konštatovať, že pomocou modelu číslo 4 sa podarilo automatickou transkripciou získať zmysluplný a zrozumiteľný text, s ktorým je možné bez väčších problémov ďalej pracovať.

a.


Civitato Insigne, Civitatis Insigne prout alteri
post 49° a sinistra e terra pro.,
minenu minenti Civitatis Neosolientis Neosoliensis
incidi curavimus Cumulum
autem Metalem ultro eunda eundo
ereximus
84 84.
Porro per jugum Jugum montis descen,,
dendo posuimus Cumulum
85
Plano ab hinc situ proceden,,
tes Cumulum erigi curavimus, curavimus
86 86.
Unde prout prius plano situ
per Iugum Montis proceden,,
do posuimus Cumulum
87
Ultro fitu situ plano procedendo
posuimus Cumulum, Cumulum
88
Leniter ab hinc per Iugum
ascendendo posuimus Cumulum
89
Lini Leni porro ascensu continua,,
to, ultimum in plaga na
Halnajowu Stalu skalu vocata po,,

b.


scopulorum saxa conspeximus, conspeximus Cumu,,
Cum lum autem Metalem posuimus
99 91
A quo jugum montis ascendendo
poscimus posuimus alium in plaga na
Kreszy Cumulum
92 92.
de Inde Iugum Montis ultro ascen,,
dendo pervenimus cum fine Diutus Ductus
ad prominentem e terra terra Lapidem
in cujus dextra parte Neosoli,,
endis ensis in sinistra vero Cremnitzi,,
ensis Civitatis insigne exsculpi
curavimus, curavimus
93.
Ab hinc situ plano euntes
posuimus Cumulum
91. 94.
Plano situ ultro quoque pro,,
cedendo, ultimum in plagana plaga na
Kreszy posuimus Cumulum
96 95
Hinc jugum montis per
Localitatem Uplaszy nomina,,
tam ascendendo posuimus
Cumulum
96

Obrázok 42 Ukážka chybovosti v overovacom súbore pri použití modelu číslo 4 (a., b.).

ZÁVER

Trénovaním vybranej vzorky rukopisného textu v platforme *Transkribus* sa postupne podarilo vytvoriť štyri modely automatickej transkripcie. Chybovosť týchto modelov na úrovni znakov v cvičnom súbore sa pohybovala v rozmedzí 1,20 – 1,71 %. Chybovosť modelov na úrovni znakov v smerodajnejšom overovacom súbore sa pohybovala v rozmedzí 4,25 – 5,35 %. Najlepším dosiahnutým výsledkom je chybovosť 4,25 % na úrovni znakov dosiahnutá pri modeli číslo 4. Vylepšenie medzi modelmi číslo 1 a 2, rovnako ako aj medzi modelmi číslo 3 a 4 sa podarilo dosiahnuť využitím základného modelu NeoLatin Ravenstein 1643 – 1772, voľne dostupného na platforme *Transkribus*. Vylepšenie medzi modelmi číslo 2 a 4 o takmer 0,5 % chybovosť na úrovni znakov sa podarilo dosiahnuť dodatočným rozšírením manuálne prepísaného textu o 10 strán a rozšírením počtu slov v cvičnom súbore o približne 2000 slov. Na základe uvedeného výsledku je možné predpokladať, že rozšírenie manuálne prepísaného textu o ďalších 10 – 15 strán a rozšírenie cvičného súboru o ďalších 2 000 – 2 500 slov by mohlo priniesť ďalšie vylepšenie modelu. Uvedené kroky by mohli viesť k vytvoreniu modelu s chybovosťou na úrovni znakov atakujúcou 4 % hranicu. Vylepšenie modelu na úroveň medzi 3 – 4 % chybovosti znakov možno považovať ako maximálne možné a dosiahnuteľné. V prípade, že sa podarí dosiahnuť takúto chybovosť na úrovni znakov, nebude potrebné už ďalej pokračovať v trénovaní modelov. Najlepší model s najnižšou mierou chybovosti na úrovni znakov potom bude možné použiť na automatickú transkripciu celého dokumentu. Dosiahnutie ešte nižšej chybovosti by bolo možné zrejme len kvalitnejším snímaním (zdigitalizovaním) dokumentu. Takéto kvalitné zosnímanie by mohlo zabezpečiť skenovanie dokumentu prostredníctvom výkonného skeneru v rozlíšení 300 dpi a viac.

Od vylepšenia modelu na úroveň 3 – 4 % chybovosti znakov zrejme netreba očakávať zásadnejšie pozitívne zmeny v identifikácii číslic, rozlišovaní malých a veľkých písmen, interpunkcie či medzier medzi slovami. V týchto aspektoch už nie je možné predpokladať zásadnejšie vylepšenie modelov automatickej transkripcie. Naopak, zníženie chybovosti na úrovni znakov pod 4 % by mohlo priniesť posun v prípade lepšej identifikácie často zamieňaných písmen. Nejaké významnejšie a výraznejšie vylepšenia v znížení chybovosti na úrovni znakov už zrejme nie je možné predpokladať.

Po vytvorení finálneho prepisu celého dokumentu metódou automatickej transkripcie bude nevyhnutné celý dokument znovu prečítať a opraviť všetky chyby. Len takýmto spôsobom bude možné dokument pripraviť na jeho eventuálne publikovanie vo forme pramennej edície. Aj bez finálnej korekcie textu však možno s textom ďalej pracovať. Takto neupravený a neskorigovaný text sa bude dať použiť napríklad na fulltextové vyhľadávanie osôb, miest či konkrétnych slov a slovných spojení uvedených v dokumente. Aj v takejto podobe bude dokument využiteľný na ďalšiu vedeckú analýzu a rozšírenie úrovne nášho poznania.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

BEL, Matej: *Zvolenská stolica*. eds. I. Nagy – M. Turóci, Čadca : Kysucké múzeum, 2017.

BUMBA, Jan: *České katastry od 11. do 21. století.* Praha : Grada, 2007.

Download Transkribus. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-10-28]. Dostupné na: https://readcoop.eu/Transkribus/download/

GRAUS, Igor: K najstaršej podobe erbu Banskej Bystrice. In: *Genealogicko-heraldický hlas*, roč. 10, č. 2, 2000, s. 16 – 22.

GRAUS, Igor: *Banská Bystrica v 16. storočí: štúdie z dejín mesta.* Banská Bystrica : Enterprise, 2015.

How To Train and Apply Handwritten Text Recognition Models in Transkribus eXpert. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-10-28]. Dostupné na: https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/

JURKOVICH, Emil: *Dejiny kráľovského mesta Banská Bystrica.* preklad I. Nagy, Banská Bystrica : Pribicer, 2005.

KATUŠČÁK, Dušan: Digital humanities a automatická transkripcia rukopisných textov. In: *ITlib: informačné technológie a knižnice*, roč. 24, č. 1, 2020, s. 6 – 16.

NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 53 – 67.

NAGY, Imrich: Transkribus ako nástroj na sprístupnenie dobových archívnych pomôcok na príklade Csákósovho katalógu korešpondencie Koháryovcov. In: *Digital humanities: nástroje sprístupňovania historického dedičstva. Zborník abstraktov*. eds. P. Maliniak – I. Nagy, Banská Bystrica : Štátna vedecká knižnica, 2022, s. 66 – 69.

NOVÁK, Jozef: *Slovenské mestské a obecné erby*. Bratislava : Slovenská archívna správa, 1967.

Slovenský národný archív v Bratislave, špecializované pracovisko Slovenský banský archív v Banskej Štiavnici, fond Banská komora v Banskej Bystrici.

Štátny archív v Banskej Bystrici, fond Mesto Banská Bystrica.

The ScanTent: Professional scanning with your smartphone. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-10-28] Dostupné na: https://readcoop.eu/scantent/

TOMEČEK, Oto: *Drevorubači a uhliari v lesoch Banskej Bystrice: k problematike osídlenia horských oblastí banskobystrického chotára do polovice 19. storočia.* Banská Bystrica : Fakulta humanitných vied UMB, 2010.

VRTEĽ, Ladislav: *Osem storočí slovenskej heraldiky.* Martin : Matica slovenská, 1999.

KAPITOLA 6

KEĎ SA STROJ UČÍ ČÍTAŤ HURBANOVE LISTY

Alica Kurhajcová
Univerzita Mateja Bela v Banskej Bystrici; Filozofická fakulta; Katedra histórie
E-mail: alica.kurhajcova@umb.sk

ABSTRAKT

Cieľom aplikovaného výskumu, ktorý kapitola predstavuje, je vytvoriť pomocou digitálneho nástroja *Transkribus* model na automatické rozpoznávanie rukopisu Jozefa Miloslava Hurbana, popredného protagonistu slovenskej politiky a kultúry 19. storočia. Opisu hlavných etáp práce a v rámci nich jednotlivých metód a postupov, ktoré sa pri tvorbe modelu nazvaného *Model J. M. Hurban* uplatnili, predchádza predstavenie zbierky Hurbanových listov. Úvodné časti ozrejmujú cestu k jej vytypovaniu, potenciál listov ako ego-dokumentov v rámci kultúrnohistorického výskumu a tiež listy ako médiá na tvorbu príslušného modelu. S ohľadom na výskumný zámer – obohatiť prostredie *Transkribu* o korpus textov jazykovo slovacikálneho charakteru – bolo dôležité vyvinúť model na Hurbanovo latinské kurzívne písmo, ktoré použil v listoch prevažne písaných v dobovej slovenčine a (biblickej) češtine. Model bol trénovaný na vybraných Hurbanových rodinných listoch, spočiatku na malých vzorkách, neskôr stredne veľkých, priebežne zdokonaľovaný a verifikovaný, v poslednej fáze dokonca (neplánovane) zmenený z technológie strojového čítania HTR+ na PyLaia. Dosiahnuté výsledky dokazujú, že stroj si vyvinul „cit“ na čítanie Hurbanovho rukopisu.

Kľúčové slová: Jozef Miloslav Hurban; osobná korešpondencia; *Transkribus*; model HTR+; PyLaia

ABSTRACT
When a machine learns to read Hurban‘s letters

The aim of the applied research presented in the chapter is to use the digital tool *Transkribus* to create a model for automatic recognition of the manuscript of Jozef Miloslav Hurban, a leading protagonist of Slovak politics and culture in the 19th century. The description of the main stages of the work and within them the individual methods and procedures that were used in the creation of the model - called the *Model J.M. Hurban* - is preceded by the presentation of the collection of Hurban’s letters. The introductory parts clarify the path to its identification, the potential of letters as ego-documents within cultural-historical research, and also letters as media for the

creation of the relevant model. With regard to the research objective – to enrich the *Transkribus* environment with a corpus of texts of a linguistically Slovak character – it was important to develop a model for Hurban's Latin cursive script, which he used in letters mostly written in contemporary Slovak and (biblical) Czech. The model was trained on selected Hurban family letters, initially on small samples, later medium-sized, continuously improved and verified, in the last phase even (unplanned) changed from HTR+ machine reading technology to PyLaia. The results achieved prove that the machine has developed a "feel" for reading Hurban's manuscript.

Key words: Jozef Miloslav Hurban; personal correspondence; *Transkribus*; model HTR+; PyLaia

ÚVOD

V súvislosti s *Transkribom*, ako na to poukazujú úvodné kapitoly tejto knihy, máme dočinenia s pomerne mladou, stále sa vyvíjajúcou softvérovou platformou, ktorá vďaka technológii strojového čítania a metóde automatického rozpoznávania rukopisných textov (HTR+, PyLaia) umožňuje spoločensko-humanitným vedcom a širšej odbornej verejnosti efektívne pracovať s tlačenými, ale predovšetkým s rukopisnými dokumentmi. Účinnosť tohto digitálneho nástroja sa zvyšuje neustálym rozširovaním platformy o nové dáta v podobe pribúdajúcich *modelov na automatickú transkripciu rukopisov*, ktoré v sebe nesú kód jednej alebo viacerých rúk, istý typ písma a jazyk(y) jednotlivca, kolektívu viacerých autorov či kultúry v rámci určitého obdobia. Konkrétne v modeli, tvorbu ktorého tu (okrem iného) predstavujem, je zakódovaný rukopis popredného protagonistu slovenského politického, cirkevného a kultúrneho života 19. storočia Jozefa Miloslava Hurbana. Pokiaľ ide o rukopis jednej osoby, musíme mať od začiatku na pamäti, že vyvíjaný model je špecifický (v kontraste k univerzálnemu), a preto jeho prínos, napr. pre pamäťové inštitúcie, treba hľadať inde ako v masovej digitalizácii a transkripcii rozličných písomností devätnásteho či iného storočia.[1] Aj keď to na úvod nevyznieva povzbudivo, výber rukopisu J. M. Hurbana nebol náhodný. V tejto kapitole sa preto popri opise hlavných etáp a metodológie, ktorú som pri tvorbe a zdokonaľovaní modelu na automatické rozpoznávanie Hurbanovho rukopisu – *Modelu J. M. Hurban* (ako som ho pracovne nazvala) uplatnila, zmienim aj o možnostiach a limitoch jeho využitia a vôbec o relevantnosti vytypovanej zbierky – *Hurbanových listov* pre súčasný kultúrnohistorický výskum.

VYTYPOVANIE ZBIERKY OSOBNEJ KOREŠPONDENCIE J. M. HURBANA

Ako bolo vyššie uvedené, výber Hurbanovho rukopisu v podobe jeho osobnej korešpondencie bol premyslený a odvíjal sa od niekoľkých kritérií. V prvom rade som vychádzala z vlastného profesijného záujmu o bádanie kultúrnych dejín dlhého devätnásteho storočia, ktorý sa spravidla nezaobíde bez štúdia ego-dokumentov, no tie neraz nachádzame len v rukopisnej

1 RABUS, Achim: Training generic models for Handwritten Text Recognition using Transkribus: Opportunities and pitfalls. In: *To appear in Proceedings of the Dark Archives Conference* [online]. Oxford, 2019, pp. 3 – 4 [cit. 2022-11-20]. Dostupné na: https://www.academia.edu/49356690/Training_generic_models_for_Handwritten_Text_Recognition_using_Transkribus_Opportunities_and_pitfalls

podobe. Práve *Transkribus* sa od spustenia projektu javil ako výzva a kľúč, ktorým by sa dala sprístupniť dosiaľ málo vydávaná (a možno aj horšie čitateľná) rukopisná pozostalosť významnej osobnosti slovenskej kultúry 19. storočia – typu denníky, korešpondencia či iné zápisky súkromnej povahy. Podmienkou zmysluplnosti vyvíjania modelu na „čítanie" rukopisu bol tiež dostatočný rozsah zbierky dokumentov a jej reprezentatívnosť (dokumentárne pokrytie veľkej časti života vybranej osobnosti). Napokon rozhodujúcim kritériom sa stala jazyková stránka textov, zmyslom čoho bolo obohatiť prostredie *Transkribu* o korpus jazykovo slovacikálnych dokumentov (najmä dobové podoby slovenčiny a v slovenskom kultúrnom prostredí aj češtiny); na začiatku projektu jestvoval len model pre rukopis Andreja Kmeťa.[2] Po zohľadnení uvedených kritérií sa v konečnom výbere ocitla neprehliadnuteľná, no stále súborne nevydaná a tým ani v dostatočnej miere nepreskúmaná *osobná korešpondencia J. M. Hurbana* z 30. až 80. rokov 19. storočia.

Čo sa týka autora listov, životný príbeh J. M. Hurbana (1817 – 1888) ani význam jeho osobnosti tu netreba zvlášť predstavovať. Dokladajú ich početné štúdie a publikácie o jeho všestrannej činnosti, niekoľko biografických prác, jubilejné čísla časopisov, edície literárnych spisov a obsiahle state v syntézach k dejinám Slovenska, slovenského povstania v rokoch 1848/49, evanjelickej cirkvi a. v. a vývinu slovenskej literatúry a spisovného jazyka, či už z pera historikov, literárnych historikov a kritikov ale aj teológov, cirkevných historikov a filozofov.[3] Vo všeobecnosti sa dá konštatovať, že Hurbanova pozícia v kontexte slovenských, širších stredoeurópskych či dokonca európskych dejín 19. storočia (napr. Hurbanovo heslo v rámci holandského projektu *Encyklopedia of Romantic Nationalism in Europe*)[4] je v súčasnosti pevne ukotvená. Ako ukazuje slovenská historiografia, nespochybniteľný je jeho vplyv a podiel na formovaní moderného slovenského národa a ideológie vrátane jednotlivých komponentov, ktoré národnú identitu v dobe moderného nacionalizmu pomáhali konštituovať (napr. spisovaný jazyk, literatúra, divadlo, cirkev, politické programy, no aj otvorený zápas proti politickým a cirkevným oponentom). V literárnohistorických prácach sa mu prisudzuje nejedna zakladateľská rola – v oblasti literárnej histórie, publicistiky a – vďaka jeho recenziám, esejam či polemickým vystúpeniam – aj literárnej kritiky a esejistiky.[5] V literárnej pamäti, čo podčiarkuje aj prax učebnicovej kanonizácie významných literárnych osobností, zostáva prozaikom slovenského literárneho romantizmu

2 KATUŠČÁK, Dušan: Digital humanities a automatická transkripcia rukopisných textov. In: *ITlib: informačné technológie a knižnice*, roč. 24, č. 1, 2020, s. 6 – 16.

3 VIRŠINSKÁ, Miriam: Bibliografia diel o J. M. Hurbanovi. In: *Jozef Miloslav Hurban – prvý predseda Slovenskej národnej rady (Príspevky k 200. výročiu narodenia)*. zost. N. Rolková Petranská, Bratislava : Kancelária NR SR, 2017, s. 240 – 250. Z novších vedeckých prác napr. KODAJOVÁ, Daniela – MACHO, Peter (eds.): *Jozef Miloslav Hurban – osobnosť v spoločnosti a reflexii*. Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 2017.

4 MACHO, Peter: Hurban, Jozef Miloslav. In: *Encyclopedia of Romantic Nationalism in Europe* [online], last changed 20.04.2022 [cit. 2002-10-30]. Dostupné na: https://ernie.uva.nl/viewer.p/21/56/object/131-158657

5 CHMEL, Rudolf: *Kritika a kontinuita*. Bratislava : Smena, 1975, s. 12 – 50; NOGE, Július: Slová, ktoré pripravovali čin. In: *Dielo I.* Bratislava : Tatran, 1983, s. 15 – 30; KRAUS, Cyril: *Začiatky slovenskej kritiky. Literárna kritika v slovenskom klasicizme a romantizme*. Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 1991, s. 117 – 208.

štúrovskej generácie.[6] Teologicko-filozofické postoje a náboženské myslenie tohto evanjelického kňaza interpretujú cirkevní historici v pohybe: od dočasného racionalistu cez prívrženca filozofického idealizmu smerom ku konzervatívnemu ortodoxnému teológovi.[7] Už za života sa Hurban stal symbolom vytrvalého zápasu, neustáleho činu a odvahy – posudzovaným oboma znamienkami: jedni k nemu vzhliadali ako k národnému hrdinovi, vodcovi, tribúnovi ľudu a bojovníkovi za práva a slobodu Slovákov, iní v ňom videli panslávskeho agitátora a rebela, vlastizradcu, reakcionára, konzervatívca a postrach spoločnosti. Aj posmrtný odborný výklad Hurbanovej osobnosti či jeho obraz v kolektívnej historickej pamäti podliehal rôznej kontextualizácii a neraz aj symbolickej a politickej inštrumentalizácii.[8] V základoch však jeho osobnosť zostala neotrasená, v istých uzlových bodoch protirečivá a zložitá.

POTENCIÁL LISTOV V RÁMCI KULTÚRNOHISTORICKÉHO VÝSKUMU

Osobnú korešpondenciu spolu s denníkmi, autobiografiami, memoármi a inými dokumentmi súkromnej proveniencie súhrnne označujeme ako *ego-dokumenty*, pojmom, ktorý v roku 1953 zaviedol holandský historik a univerzitný profesor Jacques Presser. V západoeurópskej historiografii sa aj vďaka týmto prameňom súkromnej a dôvernej povahy (aj vďaka interdisciplinárnemu dialógu a v reakcii na tradičné metódy) vytvárali v druhej polovici 20. storočia podmienky na to, aby sa „prinavrátil človek do dejín". Presserov záujem o tento typ prameňov a vôbec personalizáciu histórie podnietila marxistická interpretácia dejín, ktorá narábala s neosobnými a abstraktnými pojmami. Práve v autobiografických materiáloch nachádzal *„človeka v oveľa čistejšej a v oveľa osobnejšej forme ako v ostatných prameňoch"*[9] – namiesto bezmenného človeka sa v nich vynáralo jedinečné ego. Neskôr Presser vo svojej rozlúčkovej prednáške roku 1969 stručne charakterizoval ego-dokumenty ako pramene, v ktorých sa *„ego úmyselne alebo náhodne vynára či skrýva"*, čím podľa nemeckého historika W. Schulzeho prekročil úzke poňatie týchto dokumentov, obmedzené len na autobiografické texty.[10] Potvrdzujú to aj súčasné pohľady, že ego-dokumentmi môžu byť rôznorodé autentické individuálne záznamy (pôvodne neurčené na zverejnenie), napr. hoci aj vlastnícke záznamy v knihách, v ktorých do popredia vystupuje ego autora.[11] Je nespochybniteľné, že táto skupina prameňov sa podieľala na rozvoji sociálnych a kultúrnych dejín a v rámci nich dejín mentalít, dejín

6 KRAUS, Cyril: Próza slovenských romantikov. In: *Slovenskí romantici. Próza*. zost. C. Kraus, Bratislava : Slovenský Tatran, 2002, s. 5 – 17; BÍLIK, René: Pragmatický romantik. In: *Jozef Miloslav Hurban. Prózy a články*. Bratislava : Kalligram – Ústav slovenskej literatúry Slovenskej akadémie vied, 2014, s. 273 – 281.

7 HANUS, Radoslav: J. M. Hurban: konfesijný luteránsky teológ a politik. In: *Jozef Miloslav Hurban – prínos pre cirkev a národ. Zborník z vedecko-historickej konferencie ECAV na Slovensku.* Liptovský Mikuláš : Tranoscius, 2017, s. 24 – 47.

8 MACHO, Peter: Jozef Miloslav Hurban – náš prvý legionár? Konštruovanie kontinuity boja za národnú slobodu. In: *Revolúcia 1848/49 a historická pamäť*. Bratislava : Historický ústav SAV, s. 165 – 189.

9 BAGGERMAN, Arianne – DEKKER, Rudolf: Jacques Presser, Egodocuments and the Personal Turn in Historiography. In: *The European Journal of Life Writing*, vol. 7, 2018, pp. 90 – 91, 93.

10 SCHULZE, Winfried: Ego-Dokumente. Annäherung an den Menschen in der Geschichte? Vorüberlegungen für die Tagung „Ego-Dokumente". In: *Ego-Dokumente. Annäherung an den Menschen in der Geschichte*. zost. W. Schulze, Berlin : Akademie Verlag, 1996, s. 20.

11 KOLLÁROVÁ, Ivona: Ego-dokumenty vo výskume dejín typografického média. In: *Kniha 2014. Zborník o problémoch a dejinách knižnej kultúry*. zost. M. Domová, Martin : Slovenská národná knižnica, 2014, s. 95 – 96.

každodennosti, mikrohistórie či historickej antropológie; tento trend pokračuje dodnes,[12] pričom sme svedkami rozvíjania nových paradigiem a smerov historického bádania. Najnovšie sa na svetových fórach historikov objavuje záujem o dejiny emócií, keď sa hovorí a už aj v susedných historiografiách píše o emočnom obrate či boome.[13]

V súčasnej (aj slovenskej) historiografii sa ego-dokumenty natoľko udomácnili, že vystupujú ako svojbytné pramene a nie iba ako doplnok k prameňom inštitucionálneho charakteru. Zvýšenému záujmu o prácu s týmito osobnými dokumentmi predchádzalo už v 19. storočí, najmä v nemeckých krajinách a vo Francúzsku, ich uverejňovanie v podobe edícií prameňov. V slovenských pomeroch sa zviditeľnilo najmä vydávanie osobnej korešpondencie slovenských dejateľov 19. a prvej polovice 20. storočia (menej ich denníkov a najmä ich literárnej tvorby), k čomu však historici a zvlášť literárni historici a vedci pristúpili najskôr (a najmä) v druhej polovici minulého storočia. Z vtedajších edičných činov treba vyzdvihnúť súborne vydávané listy Slovenskou akadémiou vied v rámci edície *Korešpondencia a dokumenty*, Maticou slovenskou v edícii *Teória a výskum* (séria *Monografie – Documenta Litteraria Slovaca*), ako aj vydavateľstvom Tatran v edícii *Hviezdoslavova knižnica* (tá sa zamerala najmä na krásnu literatúru). Korešpondencia sa v nich vyhodnocuje ako dôležitý biografický a taktiež literárny dokument, ktorý podáva najautentickejšie fakty o živote, činnosti a tvorbe autora, no zároveň je aj *„výrazom a obrazom jeho vnútra“*.[14] Jej prínos však editori neposudzovali len v rovine predĺženej ruky biografického výskumu. Uvedomovali si tiež, že v autorovej dôvernej – spontánnej či vynútenej – výpovedi sa odráža aj jeho vzťah k adresátom a tým aj k systémom, v rámci ktorých sa pohybuje a identifikuje (rodina, sociálna vrstva, konfesia, región, štát).[15] Nestačí preto pristupovať ku korešpondencii len ako k svedectvám autora o sebe, jeho súkromnom, intímnom a duševnom živote. Listy neraz obsahujú aj osobné výpovede o dobových pomeroch, udalostiach a širších spoločenských javoch či súvislostiach; reflektujú dobové stereotypy či obrazy o sebe a iných. V období pred masovým rozšírením modernej periodickej tlače (v Uhorsku v poslednej tretine 19. storočia) to umocňovala aj jedna z podstatných funkcií korešpondenčného styku; totiž pre vzdelancov zo stredných a vyšších spoločenských vrstiev predstavovali listy dôležitý komunikačný kanál na sprostredkovanie nielen osobných správ, ale aj správ o záležitostiach verejných, z domu i zo zahraničia.[16] Aj preto korešpondenciu 19. storočia ocenil nejeden jej editor, a to z oboch uvedených aspektov – ako materiál, ktorý *„precizuje a skonkrétňuje naše vedomosti o autorovi v dobe i dobe v autorovi“*.[17]

Za uplynulých približne sto rokov nemožno nespozorovať, že nepomer medzi významom činorodej osobnosti a rozsahom jej vydanej korešpondencie sa azda najvýraznejšie

12 O význame ego-dokumentov (listov a denníkov mužov i žien) na začiatku 21. storočia: LENDEROVÁ, Milena – KUBEŠ, Jiří (eds.): *Osobní deník* a *korespondence – snaha o prezentaci, autoreflexi nebo (proto)literární vyjádření?* Pardubice : Univerzita Pardubice, 2004.

13 ŠVAŘÍČKOVÁ SLABÁKOVÁ, Radmila: Dějiny emocí: nové paradigma ve studiu historie. In: *Český časopis historický,* roč. 114, č. 2, 2016, s. 291 – 315.

14 ŠMATLÁK, Stanislav (ed.): *Hviezdoslav zblízka*. Bratislava : Tatran, 1985, s. 5.

15 SCHULZE, Winfried: Ego-Dokumente..., s. 28.

16 FÓNAGY, Zoltán: Levelezés a 19. századi Magyarországon. In: *Történelmi szemle*, roč. 55, č. 4, 2013, s. 629 – 630.

17 PETRUS, Pavol: Janko Jesenský v listoch. In: *Listy Janka Jesenského 1.* zost. P. Petrus, Martin : Matica slovenská, 1989, s. 5.

prejavuje u Jozefa M. Hurbana. Aj medzi vyššie menovanými edičnými projektmi sú jeho listy len nepatrne zastúpené a aj inde skôr výberovo a v rôznej kvalite spracované, resp. niekedy len prepísané bez edičnej poznámky (pokiaľ zámerom editora bolo život osobnosti populárne a čitateľsky atraktívne podať). Ako zaujímavosť možno spomenúť, že už jeho syn Svetozár Hurban Vajanský prezieravo chápal váhu súkromných dokumentov ako prameňov poznania vlastného národa, keď do *Slovenských pohľadov* roku 1906 pripravil ukážky z rodičovskej korešpondencie – listy otca Jozefa manželke Anne Hurbanovej (rodenej Jurkovičovej) – s dodatkom: *„Všetko je hodno zachovať, čo jako-tak súvisí s naším národným dielom, i keby naoko bolo dielom súkromným. Verejnosť skladá sa zo samých súkromností, jako mozaikový obraz z drobných kamienok.“*[18] *Slovenské pohľady* v podstate už od roku 1895 s prestávkami do roku 1913 uverejňovali ukážky z Hurbanovej listovej komunikácie (menovite s Jánom Hollým, Josefom V. Fričom, generálom Františkom A. Zachom, Danielom Lichardom, Andrejom Sládkovičom a Augustom H. Škultétym).[19] Okrem niekoľkých svetlých vydavateľských počinov[20] sa ukazuje, že ani zvyšné 20. storočie celkom neprialo systematickému sprístupňovaniu Hurbanových listov. Začiatkom sedemdesiatych rokov, keď R. Chmel pripravoval na vydanie Hurbanovo dielo *Slovensko a jeho život literárny*, zároveň poznamenal, že v prípade Hurbana si bude treba nevyhnutne všimnúť aj jeho korešpondenciu, ktorá nebude o nič menej časovým čítaním.[21] Všimol si ju literárny historik Tomáš Winkler. V osemdesiatych rokoch spracoval jednak biografiu J. M. Hurbana,[22] jednak popularizačno-dokumentačné dielo, v ktorom ukážkami (aj) z jeho korešpondencie Hurbanovu životnú dráhu názorne ilustroval.[23] Na kritickú edíciu si však bolo treba ešte počkať. Sľubný začiatok

18 VAJANSKÝ, Svetozár Hurban: Archiválne zlomky. In: *Slovenské pohľady*, roč. 26, č. 1, 1906, s. 1 – 7, č. 2, s. 90 – 98.

19 Jozef M. Hurban Jánovi Hollému. In: *Slovenské pohľady*, roč. 15, č. 8, 1895, s. 499 – 500; Jozef M. Hurban Jozefovi Fričovi. In: *Slovenské pohľady*, roč. 19, č. 9, 1899, s. 566 – 567; Jozef M. Hurban generálovi Zachovi. In: *Slovenské pohľady*, roč. 19, č. 9, 1899, s. 567 – 568; ŠKULTÉTY, Jozef: Smierenie J. M. Hurbana a Daniela Licharda roku 1856. In: *Slovenské pohľady*, roč. 21, č. 7, 1901, s. 393 – 405; Listy Andrejovi Sládkovičovi. In: *Slovenské pohľady*, roč. 28, č. 11 – 12, 1908, s. 763 – 764; Listy Andrejovi Sládkovičovi. In: *Slovenské pohľady*, roč. 29, č. 1, 1909, s. 57 – 58; Jozef M. Hurban Augustovi H. Škultétymu o svojej „Unii“. In: *Slovenské pohľady*, roč. 33, č. 3, 1913, s. 186 – 188.

20 KLEINSCHNITZOVÁ, Flóra: Z listov Jozefa Miloslava Hurbana Danielu Slobodovi. In: *Sborník Matice slovenskej II.* Turčiansky Sv. Martin : Matica slovenská, 1924, s. 82 – 96; *Sborník Matice slovenskej III*. Turčiansky Sv. Martin : Matica slovenská, 1925, s. 169 – 188. Z čias revolúcie je niekoľko úradných i osobných listov Hurbana uverejnených v rozsiahlych edíciách dokumentov k jednotlivým zväzkom Rapantovho monumentálneho diela *Slovenské povstanie 1848 – 49*. PETRUS, Pavol (ed.): *Korešpondencia Svetozára Hurbana Vajanského I. (Výber listov z rokov 1860 – 1890)*. Bratislava : Vydavateľstvo Slovenskej akadémie vied, 1967; ŠKVARNA, Dušan (ed.): Hurban na slovanskom juhu (júl – august 1848). In: *Od revolúcie 1848 – 1849 k dualistickému Rakúsko-Uhorsku. Pramene k dejinám Slovenska a Slovákov X*. Bratislava : LIC, 2009, s. 47 – 49; PODOLAN, Peter: Jozef Miloslav Hurban: Farníkom obce Hlboké (list). In: *Tvorba : revue pre literatúru a kultúru*, roč. 27, č. 1, 2017, s. 85 – 88; PODOLAN, Peter: List Jozefa Miloslava Hurbana Jonášovi Guothovi zo 17. mája 1851. In: *Opus tessellatum. Historia nova 14* [online]. Bratislava : Stimul, 2017, s. 95 – 99 [2022-09-10]. Dostupné na: https://fphil.uniba.sk/fileadmin/fif/katedry_pracoviska/ksd/h/Hino14.pdf; Hurbanov list Borbisovi z 18. 3. 1864 a J. Minichovi a jeho kamarátom z 18. 2. 1869: HVOŽĎARA, Miroslav: *Jozef Miloslav Hurban a jeho zápas o pravé hodnoty cirkvi a národa. Reflexia na život a prácu kňaza a národovca.* Liptovský Mikuláš : Tranoscius, 2008, s. 160 – 165, 178 – 179.

21 CHMEL, Rudolf: Hurbanovo Slovensko a jeho život literárny. In: *J. M. Hurban: Slovensko a jeho život literárny*. zost. R. Chmel, Bratislava : Tatran, 1972, s. 217.

22 WINKLER, Tomáš: *Perom a mečom*. *Biografia J. M. Hurbana.* Bratislava : Tatran, 1982.

23 WINKLER, Tomáš: *Jozef Miloslav Hurban. Život zvoniaci činom – život a dielo v dokumentoch.* Martin : Osveta, 1987.

v tomto smere odštartoval až moravský historik Zdeněk Fišer: Najskôr skromne v roku 2007, keď do piateho zväzku edície *Korespondence Aloise Vojtěcha* Šembery s názvom *Listy slovenským přátelům* zahrnul šesť Hurbanových listov adresovaných Šemberovi (od apríla 1841 do mája 1844),[24] a vo väčšom štýle a rozsahu o dva roky na to, keď štvrtý zväzok edície *Daniel Sloboda v dokumentoch* (ako súčasť väčšieho edičného projektu *Prameny dějin moravských*) venoval vzájomnej korešpondencii D. Slobodu a J. M. Hurbana. Vzhľadom na naznačený dovtedajší stav ide (podľa skromného názoru jej editora Z. Fišera) o *„oný prielom“*,[25] no podľa autora recenzie jeho diela, historika D. Škvarnu ide o *„výnimočné edičné dielo“*, okrem iného aj preto, že *„listy oboch švagrov vari výstižnejšie ako akýkoľvek iný druh dobových dokumentov kopírujú aj názorový a psychologický vývoj J. M. Hurbana. Hoci ho vcelku dobre poznáme, vnášajú doň veľa nových informácií a po ich pozornej analýze môže nadobudnúť osobnostný profil J. M. Hurbana konkrétnejší obraz.“*[26]

Hoci sa nám môže zdať, že poznanie J. M. Hurbana je nateraz vyčerpané, predsa len nevydané súkromné pramene (a to nielen listy) môžu signalizovať, že jestvujú stále neprebádané oblasti či medzery v poznaní tejto osobnosti a jej sociálnej siete. Domnievam sa, že jedným z dôvodov, ktorý môže niektorých bádateľov tohto obdobia brzdiť či odrádzať od hlbšieho ponoru do originálov jeho listov, resp. iných dosiaľ v rukopise ponechaných ego-dokumentov (napr. autobiografií, denníkových zápiskov, kázní), môže byť fyzický stav rukopisov a rôzna miera ich čitateľnosti. Uvedomujem si, že prelomiť túto bariéru pomocou inovatívneho nástroja *Transkribus*, a to vyvinutím modelu na automatické rozpoznávanie Hurbanovho rukopisu v jednotlivých etapách jeho života, je výzvou, zodpovednosťou a v kontexte rozvíjajúcich sa digitálnych humanitných vied aj novou skúsenosťou.

LISTY AKO MÉDIUM NA TVORBU MODELU V PLATFORME *TRANSKRIBUS*

Hurbanova osobná korešpondencia, ktorá bola vytypovaná na ciele aplikovaného výskumu v rámci projektu SKRIPTOR, pochádza výlučne zo zbierok a z fondov deponovaných v Literárnom archíve Slovenskej národnej knižnice v Martine: prevažne z fondu rodiny Hurbanovcov[27] a osobných fondov J. M. Hurbana a S. H. Vajanského,[28] ale aj z fondov ďalších rodín a osobností, ktoré stáli s J. M. Hurbanom v príbuzenstve (napr. rodina Jurkovičovcov či Pavol Roy, manžel Hurbanovej dcéry Boženy), v korešpondenčnom styku (napr. P. Dobšinský, A. H. Škultéty, J. Leška, J. Francisci, V. Pauliny-Tóth, Zochovci a i.) alebo

24 FIŠER, Zdeněk (ed.): *Korespondence Aloise Vojtěcha* Šembery: *Listy slovenským přátelům V.* Vysoké Mýto : Regionální muzeum, 2007.

25 FIŠER, Zdeněk: Nad korespondencí Daniela Slobody a Jozefa Miloslava Hurbana. In: *Prameny dějin moravských: Korespondence Daniela Slobody s Jozefem Miloslavem Hurbanem*. zost. Z. Fišer, Brno : Matice moravská, 2009, s. 26.

26 ŠKVARNA, Dušan: Edičný počin Zdeňka Fišera, Daniel Sloboda a Jozef Miloslav Hurban. In: *Biografické štúdie 41*. Martin : SNK – Národný biografický ústav, 2018, s. 54 – 55.

27 Literárny archív – Slovenská národná knižnica Martin (ďalej LA SNK), fond (ďalej f.) Hurbanovci, signatúra (ďalej sign.) 32. Prírastky v tomto fonde eviduje Archív z rokov 1958 – 1978. Zdroj: Mgr. Karin Šišmišová (zastupujúca vedúca oddelenia LA SNK), 8. 11. 2022 (e-mail).

28 LA SNK, f. J. M. Hurban, sign. 8 (Prírastky tohto fondu boli do Archívu prevzaté v rokoch 1955 – 1961.), sign. M 23 (Dokumenty so signatúrou začínajúcou sa písmenom M pochádzajú z Archívu SNM a boli do Literárneho archívu, vtedy Matice slovenskej, delimitované v roku 1960.). LA SNK, f. S. H. Vajanský, sign. 138 (Prírastky fondu eviduje Archív z rokov 1960 – 1974.). Zdroj: Mgr. Karin Šišmišová (zastupujúca vedúca oddelenia LA SNK), 8. 11. 2022 (e-mail).

sa iným spôsobom dostali k jeho písomnej pozostalosti (napr. literárny historik Albert Pražák). Archív nám na základe zmluvnej spolupráce medzi Univerzitou Mateja Bela a Slovenskou národnou knižnicou poskytol kvalitne vyhotovené digitálne kópie listov vo formáte TIFF (spolu 2 686 snímok), spočiatku v dvojakom rozlíšení (300 dpi a 600 dpi), zmenené v neskoršom dodatku k zmluve na rozlíšenie 600 dpi. Zbierku tvorí dohromady 902 listov z rokov 1838 – 1887, ktoré J. M. Hurban adresoval (do zbierky neboli zaradené listy neznámym adresátom, ani listy adresované právnickým osobám, napr. svetským a cirkevným úradom, inštitúciám, redakciám či spolkom):

1. rodine a príbuzným: manželke Anne (120 listov), svokrovi Samuelovi Jurkovičovi (89), synom – Svetozárovi (66), Vladimírovi (68), resp. obom (2), Bohuslavovi (33), Konštantínovi (3) a dcére Ľudmile (7), švagrom – Jurajovi Jurkovičovi (7) a Karolovi Žarnovickému (1) a švagrinej Emílii Jurkovičovej (1), neveste Ide Hurbanovej, rodenej Dobrovičovej (5), zaťom – Pavlovi Royovi (1), Dionýzovi Fejovi, manželovi dcéry Ľudmily (1), a Víťazoslavovi Lorencovi, manželovi dcéry Želmíry (1); dokopy 405 listov;
2. priateľom a blízkym spolupracovníkom, známym a (aj vysokým) predstaviteľom svetských a cirkevných úradov, inštitúcií a spolkov z Uhorska, no aj z územia a spoza hraníc habsburskej monarchie; spolu 497 listov zaslaných 151 osobám.

V dobe, z ktorej pochádzajú aj prvé Hurbanom odoslané listy, sa korešpondenčný styk obmedzoval len na úzku vrstvu spoločnosti – na jej politickú, ekonomickú a kultúrnu elitu. Až s postupujúcou spoločenskou a hospodárskou modernizáciou Uhorska v druhej polovici daného storočia (alfabetizáciou, budovaním poštovej siete, zvyšujúcou sa mobilitou v priestore a rozvojom dopravy) prestávalo byť písanie listov exkluzívnou záležitosťou elity,[29] ktorej súčasťou bol aj Hurban a jemu blízky i vzdialenejší okruh ľudí. Prevažná časť tejto korešpondencie spadá do porevolučných čias, do druhej polovice Hurbanovho aktívneho života. Jeho listy spred revolúcie, hoci štúrovská korešpondencia bola vtedy pomerne bohatá, sa zachovali len v torzovitej podobe, keďže v meruôsmom roku bola Hurbanova evanjelická fara v Hlbokom vydrancovaná a väčšina písomností zničených.[30] Komunikačný jazyk (aj písmo) volil Hurban spravidla podľa príjemcu či miesta pôsobenia, pod vplyvnom aktuálnych rečových pomerov či nám nie vždy známych vonkajších okolností: latinským kurzívnym písmom písal spravidla listy v (biblickej) češtine, dobovej slovenčine, minimálne trom príjemcom v latinčine[31] a aj jeden v maďarčine,[32] kurentom zas viacero listov v nemčine a v dvoch prípadoch azbukou v ruštine.[33] Pre náš výskumný zámer, ako už bolo v úvode spomenuté, sú zvlášť obohacujúce texty jazykovo slovacikálneho charakteru, preto som sa rozhodla vyvinúť model na *Hurbanovo latinské kurzívne* písmo. V ostatných prípadoch sa dajú použiť verejne dostupné modely umelej inteligencie na stránke READ-COOP alebo priamo v *Transkribe*,[34] napr. *German Kurrent XIX-XX M6-2* na automatický (no len približný) prepis listov písaných v kurente alebo

29 FÓNAGY, Z.: Levelezés a 19. századi Magyarországon, s. 619 – 627.

30 KODAJOVÁ, Daniela: Život s perom, krížom a mečom. In: *Jozef Miloslav Hurban – osobnosť v spoločnosti a reflexii*. Zost. D. Kodajová – P. Macho. Bratislava: VEDA, vydavateľstvo SAV, 2017, s. 22 – 23.

31 LA SNK, f. J. M. Hurban, sign. M 23 I (2, 3, 37).

32 LA SNK, f. J. M. Hurban, sign. M 23 I 29.

33 LA SNK, f. J. M. Hurban, sign. M 23 CH (24, 43).

34 *Public AI models in Transkribus*. [cit. 2022-11-22]. Dostupné na: https://readcoop.eu/transkribus/public-models/

Russian Generic Handwriting 2 na listy v azbuke. Ako príklad možno uviesť výsledok z automatického prepisu listu J. M. Hurbana písaného kurentom synovi Svetozárovi do Oberschützenu v roku 1861:[35] pri použití vyššie uvedeného modelu, vyvinutého Tobiasom Hodelom, profesorom na Univerzite v Berne, na nemecký kurent 19. storočia, sa dosiahla chybovosť 8,58 % na úrovni znakov (CER) a 27,74 % na úrovni slov (WER). Aj keď viac než osem z každých 100 znakov nebolo správne identifikovaných, prepísaný list bol pomerne dobre čitateľný a najmä zrozumiteľný.

Z vyššie predstavenej rozsiahlej zbierky som si na vycvičenie modelu vybrala (aj s ohľadom na budúce výskumné plány) listy napísané prevažne v latinke, ktoré Hurban v rokoch 1846 až 1887 adresoval najbližšej rodine – najmä manželke, deťom a čiastočne aj svokrovi. Vybrané listy, ktoré sú cennou sondou do každodenného života a vzťahov v rámci rodiny Hurbanovcov, sa ukázali ako dostatočne reprezentatívne, keďže pokrývajú pomerne dlhé obdobie Hurbanovho života, a tak zachytávajú jednotlivé odtienky vo vývine štýlu jeho rukopisu. Ako to môžeme vidieť na Hurbanovom meniacom sa rukopise (či už v závislosti od adresáta, pisateľových nálad, vnútorného rozpoloženia, vonkajších faktorov, zdravotného stavu alebo veku), výber textov jedného autora nemusí byť ešte zárukou jednotného a konzistentného rukopisu, a v tom prípade ani bezproblémovej tvorby modelu. Aj v týchto rodinných listoch sa mení jazyk: prevažne sú písané v dobovej spisovnej štúrovskej, resp. reformovanej slovenčine, inokedy v (biblickej) češtine, z času na čas obohatené o latinské sentencie, kde-tu nemecké a maďarské slová, prípadne nárečové výrazy z podbradliansko-podjavorinského kraja. Hoci do nich vstupujú aj nemecké slová či celé vety v kurente, prípadne ruština a srbčina značené azbukou, tieto typy písma som do procesu trénovania modelu nezahrnula (prakticky som s nimi nepracovala). Ako to vyzerá, keď stroj učíte prepisovať Hurbanov rukopis?

TVORBA MODELU J. M. HURBAN – POSTUPY A VÝSLEDKY

Aby mohol stroj Hurbanove listy „prečítať" – literu po litere, podmienkou je – zjednodušene povedané – naučiť ho to, čiže vycvičiť v prostredí *Transkribu* model na automatickú transkripciu Hurbanovho rukopisu, čo prebiehalo v troch, neskôr v štyroch hlavných etapách:

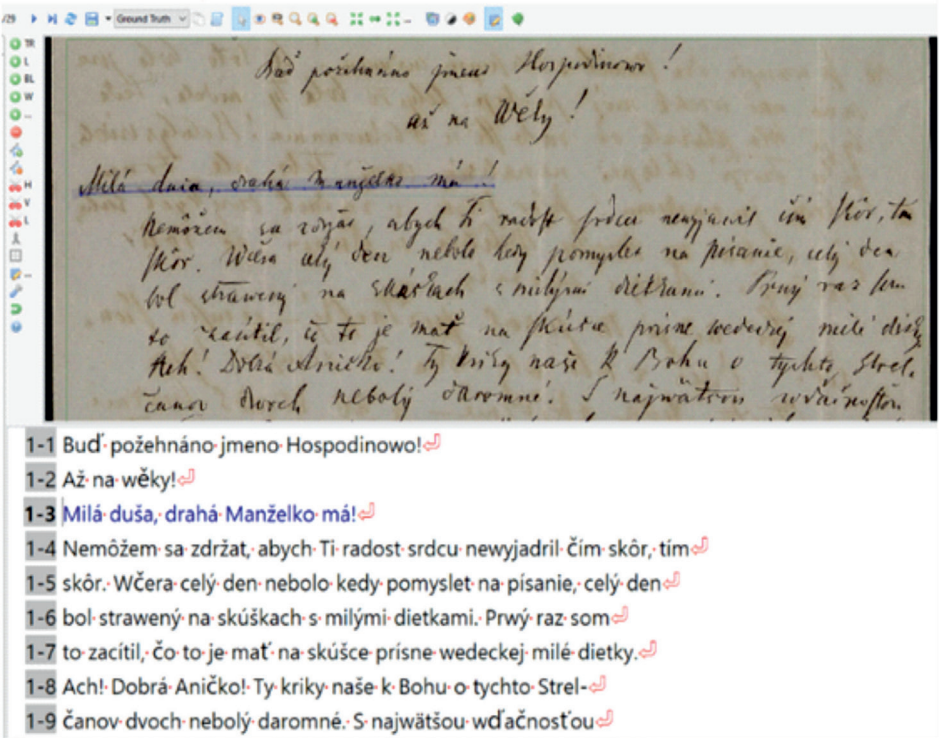
1. *Príprava a trénovanie modelu HTR+* (december 2020 – október 2021).
2. *Zdokonaľovanie modelu HTR+* (október 2021 – august 2022).
3. *Overovanie využiteľnosti modelu HTR+ v praxi* (august – október 2022).
4. *Experimentovanie s technológiou PyLaia* (november 2022).

Po oboznámení sa s princípmi fungovania platformy *Transkribus* prvú etapu charakterizovala príprava neveľkých vzoriek sčasti manuálne (dôkladne) a sčasti automaticky (strojovo) segmentovaných strán, ich vlastnoručný a čo najpresnejší prepis (príprava tzv. *Ground Truth* vzoriek) a napokon trénovanie prvých menších verzií modelu pomocou technológie HTR+ (*Handwritten Text Recognition*). Akosi intuitívne som využila A. Rabusom navrhovaný „recyklačný prístup",[36] keď sa medzi

35 LA SNK, f. Hurbanovci, List S. H. Vajanskému, sign. 32 B 56.

36 RABUS, Achim: Handwritten text recognition for croatian glagolitic. In: *Slovo* [online], roč. 72, č. 1, 2022, s. 184 – 185. [cit. 2022-10-22]. Dostupné na: https://hrcak.srce.hr/clanak/391286

prvotnými vzorkami ocitli už publikované listy Hurbana manželke. Vďaka existujúcej (hoci editorom upravenej) predlohe som tak proces vlastnoručného prepisu nielenže urýchlila, ale umožnilo mi to zároveň preniknúť do Hurbanovho rukopisu a vycvičiť sa na čítanie ďalších, ešte nevydaných listov. Na možné ťažkosti pri prepisovaní Hurbanových listov ma tiež upozornila spomenutá edícia dokumentov Z. Fišera, napr. na zamieňanie či podobnosť niektorých písmen, nedbalé používanie mäkčeňov (aj dĺžňov) a tak vytváranie novotvarov.[37]



Obrázok 43 Ukážka manuálneho prepisu originálu listu v *Transkribe*.

Počiatočné malé verzie modelu, trénované na 26 až 56 stranách (ktorých rozsah ešte nedosiahol odporúčaný minimálny počet slov – 15 000), vykazovali v overovacom súbore vysokú mieru chybovosti na úrovni znakov (od takmer 11 % až do necelých 28 % CER) a ešte väčšiu na úrovni slov (tie sú však irelevantné). Len čo sa však počet strán v cvičnom súbore takmer zdvojnásobil a zvýšil sa aj počet cyklov trénovania modelu (zo štandardných 50 na 250), výraznejšie sa znížila aj chybovosť prepisu znakov o 3 percentuálne body (*Model J. M. Hurban 28* s CER 7,98 % v overovacom súbore). Keďže ani s týmto výsledkom, zhrnutým v tabuľke 4, som sa neuspokojila (aj preto že listy v overovacom súbore nepokrývali rôzne obdobia Hurbanovho života), v druhej etape som sa zamýšľala nad ďalšími spôsobmi – ako aktuálnu skúšobnú verziu modelu vylepšiť a prepis skvalitniť.

37 FIŠER, Z.: Ediční poznámka. In: *Korespondence Daniela Slobody s Jozefem Miloslavem Hurbanem...*, s. 37.

Tabuľka 4 Prehľad trénovania prvých verzií modelu na transkripciu rukopisu J. M. Hurbana.

Dátum	Metóda	Model	Cvičný súbor		Overovací súbor	Presnosť CER %		Typ segmentácie	Počet cyklov
		ID Názov	Strany	Slová	Strany	Cvičný	Overovací		
20/07/21	CITlab HTR+	35168 Model JMH 1	26	7 673	7	0,79	15,43	M	50
21/07/21	CITlab HTR+	35187 Model JMH 2	31	8 583	2	0,96	27,64	M	50
21/07/21	CITlab HTR+	35190 Model JMH 3	28	7 936	3	0,86	15,4	M	50
22/07/21	CITlab HTR+	35217 Model JMH 4	26	2 959	3	0,62	18,9	A	50
22/07/21	CITlab HTR+	35237 Model JMH 5	28	3 189	1	0,67	18,02	A	50
22/07/21	CITlab HTR+	35 238 Model JMH 6	27	3 059	2	0,59	18,77	A	50
23/07/21	CITlab HTR+	35 257 Model JMH 7	56	11 127	6	1,78	15,75	M+A	50
23/07/21	CITlab HTR+	35 279 Model JMH 8	46	8 191	14	0,58	10,98	M+A	100
23/10/21	CITlab HTR+	37760 Model JMH 28	104	18 099	16	0,81	7,98	M+A	250

Pozn.: M – manuálna segmentácia; A – automatická segmentácia; M+A – kombinovaná segmentácia.

Osvedčili sa mi pritom dve metódy, ku ktorým spravidla dospievajú aj iní experti nástroja *Transkribus* (vrátane nášho riešiteľského tímu).[38] V prvom prípade ide o metódu priebežného zvyšovania strán v cvičnom súbore o automaticky (t. j. na základe aktuálnej verzie modelu J. M. Hurban) transkribované strany.[39] Automaticky prepísané strany sa následne opravia a pripoja k predošlým, „bezchybne" pripraveným stranám, aby sa mohlo spustiť trénovanie nového, väčšieho súboru. Tento postup som v istých časových intervaloch zopakovala šesťkrát a dopracovala som sa tak k šiestim rôzne veľkým verziám modelu (v rozpätí 123 až 560 strán v cvičnom súbore). Pri poslednej z nich – *Model J. M. Hurban 50*, trénovanej na 560 stranách (na vyše 101-tisíc slov), som dosiahla zníženie miery chybovosti na úrovni znakov o 2,08 percentuálneho bodu (zo 7,98 % na 5,9 %), ako to dokladajú aj údaje v tabuľke 5 s príslušným grafom.

38 NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika.* Roč. 51, č. 2, 2021, s. 62.

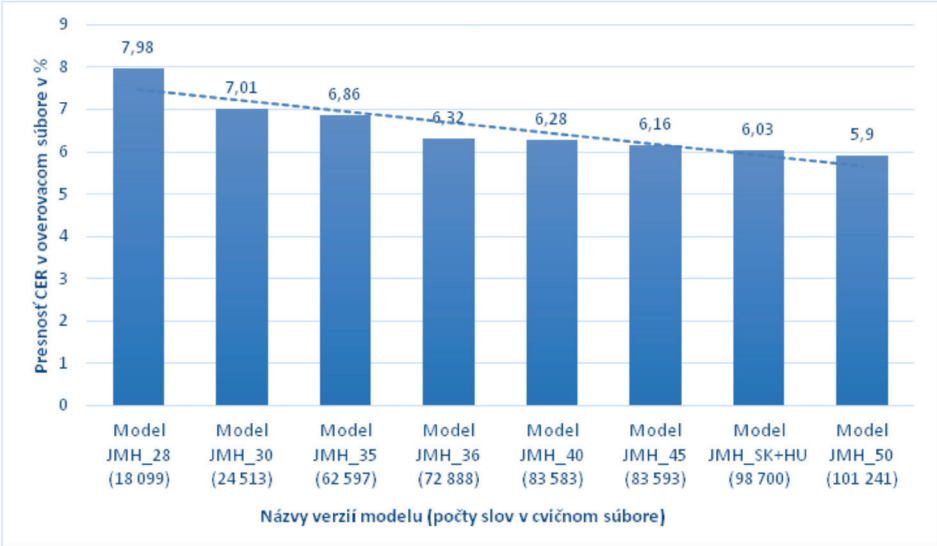
39 MASSOT, Marie-Laure – SFORZINI, Arianna – VENTRESQUE, Vincent: Transcribing Foucault's handwriting with Transkribus. In: *Journal of Data Mining and Digital Humanities*, 2019, s. 9.

Druhou efektívnou metódou, ako model zdokonaliť, je použiť v procese trénovania nového súboru tzv. základný model (*Base Model*), pričom sa tu ponúkajú dve možnosti. Základným modelom môže byť naša vlastná „najlepšia" verzia modelu, napr. pri vytváraní verzie *Modelu J. M. Hurban 30* poslúžil ako podklad *Model J. M. Hurban 28*, vďaka čomu sa miera chybovosti znížila o 0,97 percentuálneho bodu. Rovnako však môžeme použiť aj cudzí, v *Transkribe* verejne dostupný model. Pri jeho výbere je podmienkou podobnosť rukopisu a identické písmo. Keď som však ako základný model použila model na transkripciu rukopisu maďarského básnika a redaktora Józsefa Kissa z druhej polovice 19. a začiatku 20. storočia[40] (*Hungarian handwriting 19th–20th Century* s CER 9,19 %), ktorý v rámci *Transkribu* vyvinulo Centrum pre digitálne humanitné vedy Štátnej Széčéniho knižnice v Budapešti, mojím zámerom bolo ani nie tak znížiť CER hodnotu, ako skôr „zuniverzálniť" špecifický model o nové znaky a vlastnosti iného, v tomto prípade maďarského jazyka. Až na určité výkyvy, zachytené na výslednom grafe v *Transkribe*, bol konečný efekt verzie *Modelu J. M. Hurban SK+HU* uspokojivý (s CER 6,03 %).

Tabuľka 5 Prehľad trénovania a zdokonaľovania stredne veľkých verzií Modelu J. M. Hurban.

Dátum	Metóda	Model	Cvičný súbor		Overovací súbor		Presnosť CER %		Základný model HTR+	Počet cyklov
		ID Názov	Strany	Slová	Strany	Slová	Cvičný	Overovací		
23/10/21	CITlab HTR+	37760 Model JMH 28	104	18 099	16	3 341	0,81	**7,98**	-	250
27/10/21	CITlab HTR+	37807 Model JMH 30	123	24 513	16	3 341	0,79	**7,01**	*Model JMH28*	250
29/12/21	CITlab HTR+	38976 Model JMH 35	327	62 597	33	6 168	5,35	**6,86**	-	100
27/03/22	CITlab HTR+	40721 Model JMH 36	369	72 888	46	9 555	3,85	**6,32**	-	200
30/06/22	CITlab HTR+	43182 Model JMH 40	424	83 583	34	6 698	4,26	**6,28**	-	200
02/07/22	CITlab HTR+	43224 Model JMH 45	423	83 593	34	6 682	4,18	**6,16**	-	200
09/08/22	CITlab HTR+	44014 Model JMH 50	560	101 241	62	13 092	4,49	**5,9**	-	250
08/08/22	CITlab HTR+	43991 Model JMH SK+HU	530	98 700	60	12 016	4,7	**6,03**	*Hungarian handwriting 19th-20th Century*	300

40 *Kiss József levelezése a PIM Kéziratgyűjteményében* [Korešpondencia J. Kissa v rukopisných zbierkach Petőfiho literárneho múzea]. [cit. 2022-10-30]. Dostupné na: https://pim.hu/hu/digitalis-bolcseszeti-kozpont/kiss-jozsef-levelezese-pim-keziratgyujtemenyeben#



Obrázok 44 Grafické znázornenie zlepšovania Modelu J. M. Hurban.

Funkčnosť všetkých siedmich verzií modelu som overovala a porovnávala v tretej etape, a to na nových, dosiaľ netranskribovaných listoch manželov Hurbanovcov (tabuľka 6).[41] Ukázalo sa, že aj pri verzii *Modelu J. M. Hurban 50*, ktorý vykazoval najlepšie hodnoty, sa chybovosť znakov na desiatich náhodne vybraných stranách prejavila v rôznej miere: od 4,28 % (vynikajúci prepis) až do 9,3 % (použiteľný prepis) a len v jednom prípade 12,69 % (stále zrozumiteľný). Prekvapivo nižšiu mieru chybovosti (od 3,08 % do 11,17 %) na niektorých stranách som zaznamenala pri použití – ako som si ju nazvala – „univerzálnej" verzie modelu (*Model J. M. Hurban SK+HU*).

Tabuľka 6 Overenie využiteľnosti jednotlivých verzií Modelu J. M. Hurban na vybraných listoch.

	Model JMH 30 (7,01 % CER)	**Model JMH 35** (6,86 % CER)	**Model JMH 36** (6,32 % CER)	**Model JMH 40** (6,28 % CER)	**Model JMH 46** (6,16 % CER)	**Model JMH 50** (5,9 % CER)	**Model JMH SK+HU** (6,03 % CER)
	CER %	CER %	CER %	CER %	CER %	CER %	CER %
Strana 1	13,53	12,04	10,78	9,98	9,4	8,26	8,6
Strana 5	10,39	8,6	7,88	7,02	7,81	6,68	6,95
Strana 9	16,75	13,2	11,68	11,17	10,66	12,69	11,17
Strana 11	9,37	8,25	6,98	5,56	7,3	6,83	6,19
Strana 12	10,7	7,69	6,69	5,02	6,69	5,69	4,68
Strana 15	11,08	9,58	8,08	8,38	8,68	8,23	8,98
Strana 16	7,36	5,65	4,79	3,42	4,97	4,28	3,08
Strana 17	17,39	12,08	12,56	12,56	11,59	8,7	8,7
Strana 19	10,07	8,6	7,37	6,63	7,62	7,13	5,9
Strana 21	11,31	11,56	10,68	10,3	9,92	9,3	9,17
PRIEMER	**11,07**	**9,48**	**8,48**	**7,71**	**8,23**	**7,44**	**7,26**

41 LA SNK, f. Jozef Miloslav Hurban, Listy Anne Hurbanovej, sign. M23CH41, b. d.

V tomto štádiu sa potvrdili predchádzajúce výsledky a to, že optimálnymi variantmi modelu sú práve posledné dva, ktorých chybovosť na úrovni znakov (CER) v overovacom súbore predstavuje hodnotu 5,9 % (*Model J. M. Hurban 50*) a 6,03 % (*Model J. M. Hurban SK+HU*). V oboch prípadoch teda platí, že približne šesť z každých 100 znakov stroj nedokáže rozlíšiť, a teda ani správne prepísať. Aby som sa však opätovne presvedčila o účinnosti oboch týchto verzií v praxi, postup verifikácie som zopakovala na úplne novej 20-stranovej vzorke – Hurbanových väzenských listoch z Vacova (1869 – 1870).[42] Aj tu sa údaje pohybovali od špičkového prepisu (2,49 % CER, pri univerzálnej verzii modelu dokonca 2,10 % CER) až po primeraný a funkčný (8,70 % CER), iba v jednom prípade stále zrozumiteľný (13,09 % CER) (tabuľka 7).

Tabuľka 7 Overenie využiteľnosti optimálnych verzií Modelu J. M. Hurban na vybraných listoch.

	Model JMH 50		**Model JMH SK+HU**	
	CER %	WER %	CER %	WER %
Strana 1	4,4	15,58	4,75	16,08
Strana 2	7,5	26	7,72	26
Strana 3	7,42	25,45	8,7	29,9
Strana 4	8,13	26,98	9,28	29,1
Strana 5	5,48	15,74	5,55	16,33
Strana 6	5,91	21,57	5,91	25,49
Strana 7	6,64	23,08	5,63	17,95
Strana 8	8,7	26,32	7,73	22,81
Strana 10	5,53	20,45	5,3	22,73
Strana 11	7,5	24,48	8,16	25,87
Strana 12	6,42	22	5,5	18
Strana 13	2,69	10,37	3,65	11,85
Strana 14	3,61	13,33	5,09	20
Strana 15	13,09	38,24	10,74	35,29
Strana 17	3,36	13,04	4,01	14,29
Strana 19	2,49	10,49	3,54	14,51
Strana 20	2,77	11,65	**2,1**	9,22
Strana 21	2,86	11,52	3,13	13,36
Strana 22	2,8	12,14	3,28	14,08
Strana 23	3,19	11,46	4,79	18,75
PRIEMER	**4,99**	**17,54**	**5,34**	**18,84**

42 LA SNK, f. Hurbanovci, Listy Michalovi Boorovi, sign. 32 B 34; f. Jozef Miloslav Hurban, List Jozefovi Horváthovi, sign. M 23 CH 37; List Jánovi Krivošovi, sign. J 646 5; f. Hurbanovci, List Štefanovi Križanovi, sign. 32 B 64; Listy Petrovi Makovickému st., sign. M 57 A 3; List Jánovi Trokanovi, J 813; f. Albert Pražák, List Andrejovi Čajkovi, sign. 29 G 2.

Ak by vývoj v oblasti umelej inteligencie a informačných technológií nebol taký prudký, záverom by som už len optimisticky zhrnula možnosti využitia oboch verzií Modelu J. M. Hurban. Koncom roku 2022 sa však vývojársky tím *Transkribu* rozhodol HTR+ nahradiť efektívnejšou technológiou, čo viedlo k jeho pozastaveniu a tým aj znefunkčneniu existujúcich modelov HTR+. Riešením, ako nestratiť existujúce dáta, bola možnosť vo webovej platforme *Transkribus lite* automaticky pretrénovať (*retrain*) zmrazené modely HTR+ vrátane oboch verzií Modelu J. M. Hurban funkčnou technológiou PyLaia. Ďalšou možnosťou (ktorá bola k dispozícii aj predtým) bolo spustiť na základe pripravených *Ground Truth* (GT) vzoriek trénovanie nového modelu PyLaia. Aby som mohla porovnať efektívnosť transkripcie oboch typov modelov PyLaia, aj tých „nanovo trénovaných", aj tých „novovytvorených", odskúšala som v poslednej fáze tvorby modelu na Hurbanov rukopis obe ponúkajúce sa alternatívy. Hoci výsledok v prvom prípade nebol povzbudivý – po opätovnom automatickom trénovaní sa hodnota CER pri *Modeli J. M. Hurban 50* zvýšila o 2,1 percentuálneho bodu (z pôvodných 5,9 % na 8 %), kým pri *Modeli J. M. Hurban SK+HU* až o 2,67 (zo 6,03 % na 8,7 %) –, v praxi sa účinnosť prepisu pomocou oboch verzií modelu PyLaia rôznila, dokonca v niekoľkých prípadoch aj zlepšila (tabuľka 8). Naopak, miera chybovosti na úrovni znakov sa v prípade nového modelu – *Modelu J. M. Hurban PyLaia* ukázala ako optimálna (s CER 6,5 % v overovacom súbore); model bol pri nastavení 250 cyklov trénovaný na 570-stranovej GT vzorke (101 874 slov) a overovaný na 62 stranách (13 572 slov). V praktickej rovine sa však tento uspokojivý výsledok neodrazil: z troch aktuálnych verzií Modelu J. M. Hurban sa pri automatickom prepise tých istých strán prejavil *Model J. M. Hurban Pylaia* ako najhorší (v tabuľke 8 prvý stĺpec sprava). Znamená to, že hoci stroj pri tvorbe tohto modelu dokázal rozlíšiť a správne určiť necelých 94 % znakov, v skutočnosti môže byť úspešnosť na úrovni jednotlivých strán aj nižšia.

Tabuľka 8 Porovnanie CER v % pri použití pôvodných verzií modelu HTR+ a nových verzií modelu PyLaia na vzorke listov z predchádzajúcej tabuľky.

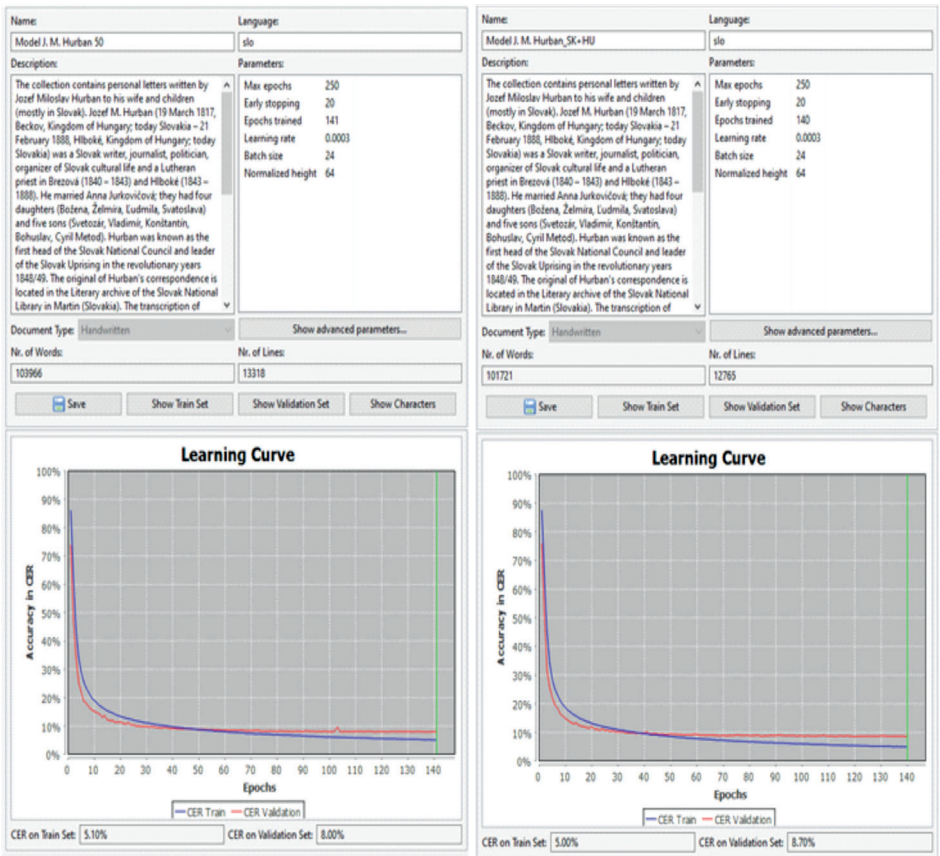
	Model JMH 50		Model JMH SK+HU		Model JMH PyLaia
	HTR+ (pôvodný)	PyLaia (RT HTR+)*	HTR+ (pôvodný)	PyLaia (RT HTR+)*	PyLaia
Strana 1	4,4	4,52	4,75	5,45	5,91
Strana 2	7,5	7,61	7,72	8,04	7,72
Strana 3	7,42	8,7	8,7	7,93	7,03
Strana 4	8,13	8,82	9,28	9,05	10,31
Strana 5	5,48	6,94	5,55	7,07	6,47
Strana 6	5,91	8,86	5,91	5,91	7,59
Strana 7	6,64	7,24	5,63	7,44	8,05

Strana 8	8,7	9,66	7,73	7,73	7,73
Strana 10	5,53	6,91	5,3	8,06	8,53
Strana 11	7,5	8,48	8,16	9,95	8,81
Strana 12	6,42	5,05	5,5	5,05	9,17
Strana 13	2,69	3,49	3,65	3,17	4,6
Strana 14	3,61	6,57	5,09	6,9	7,22
Strana 15	13,09	10,35	10,74	10,16	11,72
Strana 17	3,36	5,3	4,01	7,12	4,79
Strana 19	2,49	2,69	3,54	3,08	4,13
Strana 20	2,77	1,81	2,1	2,2	3,06
Strana 21	2,86	2,69	3,13	3,13	2,95
Strana 22	2,8	1,93	3,28	2,6	2,8
Strana 23	3,19	4,19	4,79	3,39	3,99
PRIEMER	**4,99**	**5,47**	**5,34**	**5,77**	**6,02**

*RT HTR+ (z angl. retrained Handwritten Text Recognition) – nanovo trénovaný model HTR+.

ZÁVER
„CIT“ NA ČÍTANIE J. M. HURBANA? MOŽNOSTI A LIMITY VYUŽITIA MODELU

V polčase trvania aplikovaného projektu SKRIPTOR sa mi pomocou technológie PyLaia podarilo vytvoriť dve pracovné verzie modelu na automatické rozpoznávanie rukopisu J. M. Hurbana – jeho latinského kurzívneho písma (obrázok 45). Hoci je ich chybovosť na úrovni znakov (CER) v overovacom súbore pomerne vysoká – 8 % (*Model J. M. Hurban 50*) a 8,7 % (*Model J. M. Hurban SK+HU*), možnosť pracovať na ich zlepšení podľa metód, ktoré som v rámci druhej etapy tvorby modelu navrhla (a aj uplatnila), je stále otvorená. Otázne je však, kde sa končí pomyselná hranica možnosti model zdokonaľovať. Na druhej strane k uvedeným údajom CER treba pristupovať ako k orientačným. Na výsledné hodnoty totiž neraz vplýva charakter rukopisných textov, ktoré sa počas trénovania rozhodneme vložiť do overovacieho súboru: či už ide o mieru ich reprezentatívnosti, kvalitu rukopisu alebo počet znakov na príslušnej strane (napr. v poslednom prípade platí, že čím menej znakov na strane, tým väčšie percento chybovosti). Odskúšanie oboch variantov modelu PyLaia v praxi ma presvedčilo, že presnosť prepisu a teda hodnota CER sa od listu k listu rôzni – do hry vstupujú kvalitatívne ukazovatele rukopisu, najmä ak máme dočinenia s nekonzistentným štýlom písania – od úhľadného písma až po ťažko čitateľný škrabopis. Je to aj tento prípad. Stačí, že sú niektoré písmená podobne zapísané (napr. s/z, b/t, b/l, a/u/o; k/K), prekrývajú sa či presahujú do susedného riadku, alebo sa rukopis mierne odchýli, schopnosť stroja správne identifikovať písmená a vzťahy medzi písmenami sa tým zásadne zhoršuje.



Obrázok 45 Grafické znázornenie výsledkov optimálnych verzií Modelu J. M. Hurban po ich opätovnom trénovaní technológiou PyLaia. Zdroj: *Transkribus*.

Chyby, ktorých sa stroj pri „čítaní“ Hurbanových listov dopúšťal (obrázok 46), môžeme z hľadiska kritéria zrozumiteľnosti a závažnosti rozdeliť na dve kategórie: 1. chyby zanedbateľné, resp. irelevantné (najčastejšie v interpunkcii a diakritike, pri delení slov vo vete alebo na konci riadku), 2. chyby závažné, ktoré majú vplyv na zrozumiteľnosť textu a ktoré spochybňujú účinnosť modelu (napr. chyby pri prepise číslic, skratiek, proprií či toponým, ktoré sa v prameni zriedka vyskytujú; problémy pri rozlišovaní podobných, resp. nejednoznačných znakov, ale aj viacnásobné chyby v slovách, napr. z dôvodu odklonu autora od svojho „štandardného“ štýlu písania). Overovanie prepisu tých istých listov pomocou doteraz vyvinutých verzií Modelu J. M. Hurban tiež ukázalo, že chyby sa často vyskytli v tých istých slovách, aj keď v rámci nich nie vždy v tých istých znakoch. Hoci sa tieto verzie vo výsledkoch transkripcie zásadnejšie nerozchádzajú, mienim využívať v budúcnosti aj (v poradí) druhú najlepšiu verziu (*Model J. M. Hurban SK+HU*) – nielen preto, že v niektorých prípadoch vykázala lepšiu mieru rozpoznávania znakov, ale aj preto že pri jej použití na Hurbanov rukopis možno predpokladať zohľadňovanie aj špecifických vlastností maďarského jazyka.

Blahorodný! Blahorodny Pane!
Už na slobode súc odpowiedom odpowiedam na Váš ctený ctený
dopis zo dňa 22 aprila, s ktorým ste mi
i diplom čestného mešťanstwa Vášho mesta
poslať račili. Prijmite Prijmite odo mňa srdečnú
vďaku za toto wyznačenie mna mňa vtody, keď
som odsúdený sedel vo wäzení wäzení a powedzte
mojím pánom spolumešťanom, že wyznačenie
toto horkosti môjho wäzenia značne osladilo osladilo.
Mimo týchto osobných citov wzbudili ste, pani páni
moji vo mne ale tiež aj iné radostné a bla-
žiace mňa utechy, útechy, minulý čas wyoral w srdci
mojom hlboké brázdy zármutkov všelijakých všelijakých,
ale tye sú už za mnou, tye pomohol pre-
niesť Pán, a na tye sa ja už neohliadam
späť späť, ale hladím len k cielu, ktorý wytknul
som si za raná rana mladosti mojej, a k tomuto cielu
bežiac vidím s potešením velikým jako mi
pribýva spolubojowníkow.
Činom tím, že ste mňa, tolko už honeného
a prenasledovaného pričislili pričislili k mešťanstwu
Tvojmu svojmu dokázali ste, že náležíte k tým,
čo boria sa už od dávna, predávna spolu so
mnou za slobodu rowne právo, slobodu, rowné právo, rozkvet
a blahobyt duchovný i hmotný národa slo-
venského. Lebo môžu už hovoriť čo chcejú

rite týchto národow!? tychto národow!!! Sláwa
Slovenskému Národu!
Váš! Váš, Bratia moji,
tu w 6bti 66tý den fotografowaný
Dr Hurban
Pané Dané vo Väsení Väzení
dňa 13 febr. (w 70t. hen 70ty den
wäzenia môjho) 1876 mojho) 1870.
k Prikladam NB Prikladám jednu fotografiu
pre brata professoru professora w Mikuláši Mikuláši.
Ho NB Prikladám ešte jednu fotografiu
moju pre milých Swätosatiuských SwätoMartinských
mojich priatelow a priatelkyne priatelkyne k ni ru-
Vám kám Presidentky Živeny, aniu činim
to preto pri tejto príležitosti, že už sme
skusili skúsili tu vo foznazi fogházi, ako moji moje do
Zarea ponelky hi lapane Prosiu Turca posielky sú lapané. Prosím odoslat
to ta - keď i bude u vileżito Príležitost.

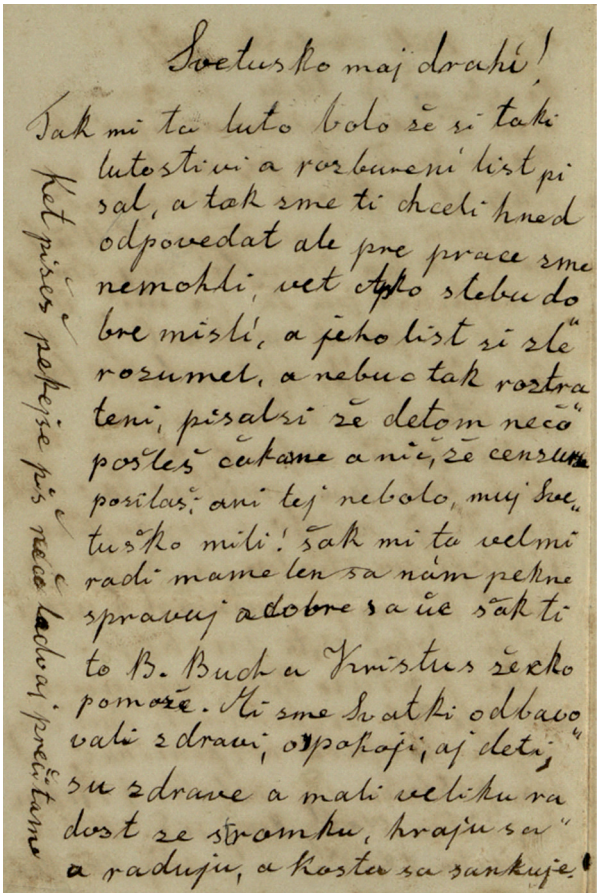
Obrázok 46 Ukážky automaticky transkribovaných strán verziami modelu PyLaia: najlepší (CER 1,81 %) a najhorší (CER 11,72 %) prepis. Zdroj: *Transkribus*.

Limity stroja pri takomto špecifickom modeli rukopisu sa prejavujú v tom, že na cudziu ruku je nepoužiteľný. Napríklad úhľadný rukopis Anny Hurbanovej (obrázok 47) ľahšie a rýchlejšie prepíšeme vlastnoručne ako s použitím aktuálnych verzií modelu PyLaia, po ktorých by bolo treba opraviť takmer 1/5 textu (čiže takmer 20 % CER). Otázna je tiež jeho použiteľnosť pri Hurbanových raných rukopisoch z 30. rokov, vzorky ktorých sa do trénovania nedostali z dôvodu malého počtu listov z daného obdobia. Stojí však za zamyslenie, či by sa v budúcnosti nedal vytvoriť univerzálny model na rukopisy osobností slovenskej kultúry 19. storočia za podmienky, žeby sa do cvičnej vzorky pridalo viacero ich rozličných autografov.

Dosiahnuté výsledky sú nateraz dostatočným dôkazom, že stroj si vyvinul „cit“ na čítanie Hurbanovho rukopisu – bezchybne dokázal prepísať dokonca aj niektoré ťažko čitateľné slová, hoci paradoxne pri niektorých pre človeka ľahko čitateľných slovách sa zas dopúšťal drobných omylov. Vyvinuté verzie Modelu J. M. Hurban sa aj v tejto podobe, po ich zverejnení, dajú v prostredí *Transkribu* rozlične využiť: či už pri štúdiu a analýze Hurbanových rukopisných textov[43] a vyhľadávaní kľúčových slov (a to aj pri ich len približnom prepise), alebo na prípravu knižnej alebo digitálnej komentovanej edície Hurbanovej korešpondencie (resp. jej časti z hľadiska tém, adresátov a pod.). Vytypované listy sú cennou osobnou výpoveďou tak o samom autorovi, ako aj o dobových skutočnostiach v tom najširšom zmysle slova. Môžu pomôcť precizovať, ako vnímal či prežíval aktuálnu realitu, aké mal názory a predstavy o blízkom i vzdialenom svete, pozemskom i transcendentálnom živote; umožňujú tiež objasniť meniace sa postoje a motívy jeho konania v konkrétnych situáciách, hlbšie preniknúť do vzťahov

43 V LA SNK sa v Hurbanovej osobnej pozostalosti okrem korešpondencie nachádzajú aj jeho denníkové záznamy, zápisníky a poznámky zo 40. až 80. rokov a vlastné životopisy. Hurbanove listy sú rozptýlené aj v archívoch cirkevných a pamäťových inštitúcií na Slovensku i za hranicami (napr. Ústredný archív ECAV v Bratislave a archívy cirkevných zborov ECAV na Slovensku; Literární archiv – Památník národního písemnictví a Archiv Národního muzea v Prahe). Ďalšie Hurbanove rukopisy sa nachádzajú v zbierke historickej knižnice Tranoscia v Liptovskom Mikuláši (pod správou Múzea Janka Kráľa v L. Mikuláši) a v Lyceálnej knižnici Ústrednej knižnice SAV v Bratislave.

s komunikujúcimi stranami, ako aj odhaliť autorove obavy či emocionálne rozpoloženie v závislosti od verejného diania či udalostí v súkromnej (rodinnej) sfére. Zároveň sú sondou do fungovania spoločnosti 19. storočia – do jej každodennej praxe (rodina, bývanie, stravovanie, starostlivosť o zdravie, cestovanie, výchova a vzdelávanie, čítanie a i.) i mentálneho sveta (diškurz, predstavy, obrazy, stereotypy).



Obrázok 47 Ukážka rukopisu Anny Hurbanovej. Zdroj: LA SNK, sign. 138 M 1.

Poďakovanie

Poďakovanie za vyhotovenie a poskytnutie digitálnych kópií Hurbanových listov vrátane informácií k fondom a zbierkam patrí Literárnemu archívu SNK v Martine, menovite zastupujúcej vedúcej oddelenia Mgr. Karin Šišmišovej a pracovníkovi archívu Mgr. Jánovi Olexovi. Za dôkladné manuálne segmentovanie prvých vzoriek z vytypovanej zbierky Hurbanových listov vďačí autorka riaditeľke Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici Mgr. Michaele Mikuškovej a vedúcej oddelenia podpory vedy Mgr. Lucii Nižníkovej, ktoré sú zároveň riešiteľkami projektu SKRIPTOR.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

BAGGERMAN, Arianne – DEKKER, Rudolf: Jacques Presser, Egodocuments and the Personal Turn in Historiography. In: *The European Journal of Life Writing*, vol. 7, 2018, pp. 90 – 110.

BÍLIK, René: Pragmatický romantik. In: *Jozef Miloslav Hurban. Prózy a články*. Bratislava : Kalligram – Ústav slovenskej literatúry Slovenskej akadémie vied, 2014, s. 273 – 281.

FIŠER, Zdeněk (ed.): *Korespondence Aloise Vojtěcha Šembery: Listy slovenským přátelům V.* Vysoké Mýto : Regionální muzeum, 2007.

FIŠER, Zdeněk (ed.): *Prameny dějin moravských: Korespondence Daniela Slobody s Jozefem Miloslavem Hurbanem.* Brno : Matice moravská, 2009.

FÓNAGY, Zoltán: Levelezés a 19. századi Magyarországon. In: *Történelmi szemle*, roč. 55, č. 4, 2013, s. 619 – 636.

HANUS, Radoslav: J. M. Hurban: konfesijný luteránsky teológ a politik. In: *Jozef Miloslav Hurban – prínos pre cirkev a národ. Zborník z vedecko-historickej konferencie ECAV na Slovensku.* Liptovský Mikuláš : Tranoscius, 2017, s. 24 – 47.

HVOŽĎARA, Miroslav: *Jozef Miloslav Hurban a jeho zápas o pravé hodnoty cirkvi a národa. Reflexia na život a prácu kňaza a národovca*. Liptovský Mikuláš : Tranoscius, 2008.

CHMEL, Rudolf: *Kritika a kontinuita*. Bratislava : Smena, 1975.

CHMEL, Rudolf: Hurbanovo Slovensko a jeho život literárny. In: *J. M. Hurban: Slovensko a jeho život literárny.* zost. R. Chmel, Bratislava : Tatran, 1972, s. 215 – 230.

KATUŠČÁK, Dušan: Digital humanities a automatická transkripcia rukopisných textov. In: *ITlib: informačné technológie a knižnice*, roč. 24, č. 1, 2020, s. 6 – 16.

Kiss József levelezése a PIM Kéziratgyűjteményében. In: *Digitális Bölcsészeti Központ* [online]. Budapest : Petőfi Irodalmi Múzeum [cit. 2022-10-30]. Dostupné na: https://pim.hu/hu/digitalis-bolcseszeti-kozpont/kiss-jozsef-levelezese-pim-keziratgyujtemenyeben#

KLEINSCHNITZOVÁ, Flóra: Z listov Jozefa Miloslava Hurbana Danielu Slobodovi. In: *Sborník Matice slovenskej II.* Turčiansky Sv. Martin : Matica slovenská, 1924, s. 82 – 96.

KLEINSCHNITZOVÁ, Flóra: Z listov Jozefa Miloslava Hurbana Danielu Slobodovi. In: *Sborník Matice slovenskej III.* Turčiansky Sv. Martin : Matica slovenská, 1925, s. 169 – 188.

KODAJOVÁ, Daniela – MACHO, Peter (eds.): *Jozef Miloslav Hurban – osobnosť v spoločnosti a reflexii*. Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 2017.

KODAJOVÁ, Daniela: Život s perom, krížom a mečom. In: *Jozef Miloslav Hurban – osobnosť v spoločnosti a reflexii*. zost. D. Kodajová – P. Macho, Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied, 2017, s. 18 – 38.

KOLLÁROVÁ Ivona: Ego-dokumenty vo výskume dejín typografického média. In: *Kniha 2014. Zborník o problémoch a dejinách knižnej kultúry.* zost. M. Domová, Martin : Slovenská národná knižnica, 2014, s. 95 – 105.

KRAUS, Cyril: *Začiatky slovenskej kritiky. Literárna kritika v slovenskom klasicizme a romantizme*. Bratislava : Veda, vydavateľstvo Slovenskej akadémie vied, 1991.

KRAUS, Cyril (ed.): *Slovenskí romantici. Próza.* Bratislava : Slovenský Tatran, 2002.

LENDEROVÁ, Milena – KUBEŠ, Jiří (eds.): *Osobní deník a korespondence – snaha o prezentaci, autoreflexi nebo (proto)literární vyjádření?* Pardubice : Univerzita Pardubice, 2004.

Literárny archív – Slovenská národná knižnica Martin, fond J. M. Hurbana (sign. 8, sign. M 23), Hurbanovcov (sign. 32), S. H. Vajanského (sign. 138).

MACHO, Peter: Hurban, Jozef Miloslav. In: *Encyclopedia of Romantic Nationalism in Europe* [online], last changed 20.04.2022 [cit. 2022-10-30]. Dostupné na: https://ernie.uva.nl/viewer.p/21/56/object/131-158657

MACHO, Peter: Jozef Miloslav Hurban – náš prvý legionár? Konštruovanie kontinuity boja za národnú slobodu. In: *Revolúcia 1848/49 a historická pamäť*. Bratislava : Historický ústav Slovenskej akadémie vied, 2012, s. 165 – 189.

MASSOT, Marie-Laure – SFORZINI, Arianna – VENTRESQUE, Vincent: Transcribing Foucault's handwriting with Transkribus. In: *Journal of Data Mining and Digital Humanities*, 2019, s. 1 – 17.

NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 53 – 67.

NOGE, Július: Slová, ktoré pripravovali čin. In: *Dielo I.* Bratislava : Tatran, 1983, s. 15 – 30.

PETRUS, Pavol: Janko Jesenský v listoch. In: *Listy Janka Jesenského 1.* zost. P. Petrus, Martin : Matica slovenská,1989, s. 5 – 16.

PETRUS, Pavol (ed.): *Korešpondencia Svetozára Hurbana Vajanského I. (Výber listov z rokov 1860 – 1890).* Bratislava : Vydavateľstvo Slovenskej akadémie vied, 1967.

PODOLAN, Peter: Jozef Miloslav Hurban: Farníkom obce Hlboké (list). In: *Tvorba: revue pre literatúru a kultúru*, roč. 27, č. 1, 2017, s. 85 – 88.

PODOLAN, Peter: List Jozefa Miloslava Hurbana Jonášovi Guothovi zo 17. mája 1851. In: *Opus tessellatum. Historia nova 14* [online]. Bratislava : Stimul, 2017, s. 95 – 99 [cit. 2022-09-10]. Dostupné na: https://fphil.uniba.sk/fileadmin/fif/katedry_pracoviska/ksd/h/Hino14.pdf

Public AI models in Transkribus. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-11-22]. Dostupné na: https://readcoop.eu/Transkribus/public-models/

RABUS, Achim: Training generic models for Handwritten Text Recognition using

Transkribus: Opportunities and pitfalls. In: *To appear in Proceedings of the Dark Archives Conference* [online]. Oxford, 2019, pp. 1 – 14. [cit. 2022-11-20]. Dostupné na: https://www.academia.edu/49356690/Training_generic_models_for_Handwritten_Text_Recognition_using_Transkribus_Opportunities_and_pitfalls

RABUS, Achim: Handwritten text recognition for croatian glagolitic. In: *Slovo* [online], roč. 72, č. 1, 2022, s. 184 – 185 [cit. 2022-10-20]. Dostupné na: https://hrcak.srce.hr/clanak/391286

SCHULZE, Winfried: Ego-Dokumente. Annäherung an den Menschen in der Geschichte? Vorüberlegungen für die Tagung „Ego-Dokumente". In: *Ego-Dokumente. Annäherung an den Menschen in der Geschichte.* zost. W. Schulze, Berlin : Akademie Verlag, 1996, s. 11 – 30.

ŠKVARNA, Dušan: Edičný počin Zdeňka Fišera, Daniel Sloboda a Jozef Miloslav Hurban. In: *Biografické štúdie 41.* Martin : Slovenská národná knižnica – Národný biografický ústav, 2018, s. 50 – 58.

ŠKVARNA, Dušan (ed.): Hurban na slovanskom juhu (júl – august 1848). In: *Od revolúcie 1848 – 1849 k dualistickému Rakúsko-Uhorsku. Pramene k dejinám Slovenska a Slovákov X.* Bratislava : LIC, 2009, s. 47 – 49.

ŠMATLÁK, Stanislav (ed.): *Hviezdoslav zblízka.* Bratislava : Tatran, 1985.

ŠVAŘÍČKOVÁ SLABÁKOVÁ, Radmila: Dějiny emocí: nové paradigma ve studiu historie. In: *Český časopis historický,* roč. 114, č. 2, 2016, s. 291 – 315.

VIRŠINSKÁ, Miriam: Bibliografia diel o J. M. Hurbanovi. In: *Jozef Miloslav Hurban – prvý predseda Slovenskej národnej rady (Príspevky k 200. výročiu narodenia)*. zost. N. Rolková Petranská, Bratislava : Kancelária NR SR, 2017, s. 240 – 250.

WINKLER, Tomáš: *Jozef Miloslav Hurban. Život zvoniaci činom – život a dielo v dokumentoch.* Martin : Osveta, 1987.

WINKLER, Tomáš: *Perom a mečom. Biografia J. M. Hurbana.* Bratislava : Tatran, 1982.

KAPITOLA 7

APLIKÁCIA AUTOMATICKEJ TRANSKRIPCIE NA PRÍKLADE KNIŽNIČNÉHO KATALÓGU ZO ZAČIATKU 19. STOROČIA

Mária Bôbová
Štátna vedecká knižnica v Banskej Bystrici
E-mail: maria.bobova@svkbb.eu

ABSTRAKT

Teologický seminár sv. Karola Boromejského v Banskej Bystrici vznikol začiatkom 19. storočia s cieľom vychovávať kňazov. V procese výchovy seminaristov zohrávala významnú úlohu seminárna knižnica. Ručne písaný dokument *Elenchus librorum* vytvorený v 19. storočí, je významným prameňom štúdia histórie knižnice. Približuje prvé obdobie jej činnosti. Poskytuje vynikajúci zdroj informácií o obsahu knižnice a systémoch organizácie. Príspevok predstaví proces tvorby modelu rukopisu pomocou platformy *Transkribus*. Hlavný dôraz sa kladie na preskúmanie možností a postupov automatickej transkripcie dokumentu skladajúceho sa z viacerých typov písma.

Kľúčové slová: automatická transkripcia, model HTR+, latinčina, nemčina, Banská Bystrica, historické knižnice, knižničné katalógy

ABSTRACT

Application of automatic transcription on the example of a **library catalog from the beginning of the 19th century**

The Theological Seminary of St. Karol Boromejský in Banská Bystrica was established at the beginning of the 19th century with an aim to educate the priests. The Seminary Library played an important role in the process of educating seminarians. The hand-written document *Elenchus librorum* created in the 19th century is a significant source for studying the history of the library. It illustrates the first period of the activities of the library. It also provides an excellent source of information about the content of the library and the systems of organization. The contribution will present the creation process of a manuscript model using the *Transkribus* platform. The main emphasis is placed on examining the possibilities and procedures of automatic transcription of document consisting of several font types.

Keywords: automatic transcription, HTR+ Model, Latin language, German language, Banská Bystrica, historical libraries, library catalogues

ÚVOD

Dejiny knižnej kultúry sú jednou z oblastí, v ktorej sa dajú naplno uplatniť aktuálne trendy digitálnych nástrojov a technológií. Tematickým príkladom sú historické knižnice, ktoré predstavujú bohatý pramenný artefakt duchovného i hmotného vývoja našej minulosti najmä vo vzťahu k vzdelávaniu, kultúre a výchove. Jednou z nich je knižnica Kňazského seminára sv. Karola Boromejského v Banskej Bystrici. Zdrojom dôležitých informácií je jej zachovaný rukopisný katalóg z 19. storočia, v ktorom možno aplikovať niekoľko funkcionalít platformy *Transkribus*, pričom prioritou je realizácia jeho automatickej transkripcie. Získané fakty sa stanú podkladom pre zodpovedanie viacerých nedoriešených otázok počiatkov knižnice i seminára.

KŇAZSKÝ SEMINÁR SV. KAROLA BOROMEJSKÉHO V BANSKEJ BYSTRICI A JEHO KNIŽNICA

Kňazský seminár sv. Karola Boromejského začal pôsobiť v Banskej Bystrici síce až začiatkom 19. storočia, ale jeho vznik a fungovanie bolo nepriamo ovplyvnené aj dianím v druhej polovici 18. storočia. V roku 1773 vyhlásením panovníčky Márie Terézie vstúpila aj v Uhorsku do platnosti bula pápeža Klementa XIV. *Dominus ac Redemptor* o zrušení jezuitskej rehole. Ich majetok bol prevedený do študijného fondu, čo malo neskôr zásadný vplyv na fond budúcej seminárnej knižnice. V roku 1776 odčlenením od Ostrihomskej arcidiecézy sa zriadili tri nové biskupstvá, medzi nimi bolo vytvorené i banskobystrické so sídlom v Banskej Bystrici. Otázka jej kňazského dorastu sa začala riešiť už v tomto období, ale spočiatku len prostredníctvom štúdií v už existujúcich kňazských seminároch.

Od roku 1800 sa stal v poradí druhým biskupom novej diecézy Gabriel Serdaheli (1742 – 1813), ktorý značné úsilie venoval na vybudovanie vzdelávacej inštitúcie pre kňazov diecézy na jej vlastnej pôde. Aj vďaka jeho osobnej angažovanosti panovník František I. Habsburský v roku 1802 na kráľovskom sneme povolil budovať semináre v biskupských rezidenciách. V roku 1804 osobitným prípisom Miestodržiteľskej rady v Budíne bolo pre Banskú Bystricu ako budúce sídlo kňazského seminára pridelených 33 diecéznych alumnistov. Následne v rokoch 1805 až 1807 bol seminár s patrocíniom sv. Karola Boromejského vybudovaný.[1]

Na činnosť seminára v jeho počiatočnom období vplývala aj reorganizácia škôl, ktorá prebiehala v Uhorsku koncom 18. a začiatkom 19. storočia a položila základy novovekého školského systému. Školstvo sa stalo prvoradou súčasťou kultúrnej politiky. V dokumente *Ratio educationis* z roku 1806 sa teologické výchovné ústavy založené diecéznymi biskupmi označovali pojmom biskupské lýceá. Ich správa bola zverená biskupom, ktorí sa mali riadiť predpismi a nariadeniami poriadku v otázkach výchovy a vzdelávania klerikov, ako i v otázkach ich profesorov a predstavených. Platil pre ne navrhnutý teologickú kurz pre teologické štúdium ako na kráľovskej univerzite. Boli im určené aj pravidlá výučby a predpísané predmety, medzi ktoré patrili Sväté písmo, cirkevné dejiny, teologické prednášky, dogmatická teológia, morálna teológia, cirkevné

1 KONIAROVÁ, Anna: *Dejiny banskobystrickej diecézy v 18. a 19. storočí.* Badín : Kňazský seminár sv. Františka Xaverského, 2002, s. 145 – 146.

právo, ďalej kazateľstvo, metóda vyučovania katechizmu, praktické vzdelávanie vo svätej liturgii a v cirkevných obradoch. Stanovené boli aj vhodné učebnice schválené panovníkom.[2]

Vyučovanie v banskobystrickom seminári sa aj na základe týchto podmienok začalo v roku 1807. Jeho prvými predstavenými boli: prefekt František Molnár, viceprefekt František Lašovský a študijný prefekt Peter Kern. V prvých rokoch sa v seminári prednášali len odborné teoretické teologické disciplíny. Zásluhou biskupa Jozefa Belánskeho v roku 1825 došlo k zavedeniu uhorského práva ako prvého predmetu mimoteologickej povahy. Neskôr sa rozvrh obohatil aj o neteologické predmety. Tiež v rámci samovzdelávacích krúžkov sa seminaristi stretávali so základmi odborov, ktoré mali využívať aj vo svojej ďalšej pastorálnej praxi. Medzi také patrili napr. pravidlá zdvorilosti, vedenie kancelárie, poľnohospodárstvo, včelárstvo a pod.[3]

Kňazské semináre ako inštitúcie zamerané na vzdelávanie vytvárali pre potreby študujúcich a prednášajúcich knižnice, ktoré sa stávali súčasťou inštitúcie a zdieľali s ňou jej osudy. Prejavovali sa tu obdobia ich slávy i obdobia ich úpadku. Preto možno knižnice brať ako prameň ich výskumu, najmä keď sú podložené písomnými materiálmi, akým je i rukopisný katalóg knižnice.

Kňazský seminár v Banskej Bystrici mal tiež svoju knižnicu budovanú od počiatku seminára. Aj tu vidíme vplyv zakladateľa seminára biskupa Serdaheliho, ktorý videl jej dôležitosť a postaral sa aj o jej vybudovanie.

S knižnicou sa spája niekoľko nevyjasnených otázok. Prvou je problematické označenie presného dátumu jej vzniku, pretože v odbornej literatúre sa stretávame s viacerými údajmi. Najstarším spomedzi nich je rok 1802, príp. 1803. Tento dátum uvádzajú uhorské štatistiky z konca 19. a začiatku 20. storočia zrejme na základe podkladov samotného seminára.[4] Súvisí s obdobím počiatočných snáh o vznik inštitúcie. Za pravdepodobný môžeme označiť aj rok 1807 ako rok, keď sa začalo vyučovať v seminári. Tento nesúlad v otázke vzniku knižnice kladie otázky aj ohľadom vzniku a vytvorenia prvého súpisu jej kníh, ktorý je nedatovaný. Bádatelia v minulosti, napr. historik Vendelín Jankovič, udávali rok 1835, zrejme pod vplyvom niektorých údajov z rukopisného katalógu.[5] Iný materiál udáva rok 1837 zrejme podľa údajov výpožičného denníka katalógu.[6]

Najstaršie písomné záznamy o seminárnej knižnici poskytuje kanonická vizitácia biskupa Belánskeho z roku 1829.[7] Zaznamenáva len zbežné informácie o jej fonde, fungovaní a umiestnení.

2 *Ratio educationis 1777 a 1806: prvá jednotná sústava výchovy a vzdelávania v dejinách našej kultúry*. Bratislava : Slovenské pedagogické nakladateľstvo, 1988, s. 326 – 330.

3 MIŠÍK, Mikuláš: Banskobystrickí bohoslovci v slovenskom národnom živote. In: *Zborník literárno-vedeckého odboru Spolku sv. Vojtecha*, tom. 2, vol. 2, 1935, s. 9 – 153.

4 GYÖRGY, Aladár (ed.): *Hivatalos statisztikai közlemények: magyarsrszág köz- és magánkönytárai 1885-ben*. Budapest : Athenaeum, 1886, s. 276.

5 JANKOVIČ, Vendelín: *Rukopisy Miestneho pracoviska Matice slovenskej v Bratislave*. Martin : Matica slovenská, 1958, s. 143.

6 *Zoznam rukopisných katalógov historických knižníc na Slovensku*. Martin : Knihovedné stredisko MS, 1959.

7 Diecézny archív v Banskej Bystrici, fond Kanonické vizitácie, sign. CV 24: Visitatio canonica Josephi Belanszky, episcopi Neosoliensis, anno MDCCCXXIX, s. 392 – 393.

Podľa nich bol fond knižnice v jej počiatočnom období zostavený z viacerých častí. V prvom rade ho tvorili knihy zo zrušenej jezuitskej knižnice v neupresnenom počte. Môžeme však vychádzať z existujúcich katalógov tejto knižnice v čase zrušenia rádu, ktoré uvádzajú vyše 4 000 kníh so širokým záberom vedných disciplín. Následná likvidácia a znovurozdelenie jezuitského knižného materiálu prebiehali v rokoch 1773 – 1790, pričom časť kníh bolo odvezených do Budapešti a Košíc. Zvyšná časť podľa nariadenia Miestodržiteľskej rady z roku 1790 zostala v úschove miestneho katolíckeho gymnázia a stala sa základom pre viaceré miestne knižnice, gymnaziálnu, kapitulskú a seminárnu knižnicu.[8] Seminárnu knižnicu obohatili aj knihy z farností diecézy. Už počas kanonickej vizitácie kapituly a diecézy, ktorá prebiehala v rokoch 1802 až 1805, žiadal biskup Serdaheli označovať nevyužité knihy vo farských knižniciach s úmyslom ich neskoršieho zaradenia do seminárnej knižnice. Ďalšie knihy darovali z vlastných knižníc biskup Serdaheli, Ľudovít Benický a niekoľko učebníc aj trnavský seminár.

Takto sa vytvoril počiatočný fond knižnice, ktorý podľa údajov prvého katalógu obsahoval spolu okolo 3 600 zväzkov. Už v úvode svojho pôsobenia sa seminár mohol spoľahnúť na dobre vybavenú knižnicu, v ktorej prevládali diela teologického charakteru, ale zastúpenie mali v nej aj humanitné disciplíny, ako dejiny, právo, filozofia, jazykoveda, pedagogika a ďalšie. Fond sa postupne rozširoval prostredníctvom darov prevažne duchovných osôb. Poskytli ich napríklad Tomáš Červeň, Imrich Vály, Dominik Benedict, Martin Kiseli[9]. Spravidla išlo o menšie počty kníh. Len v prípade pozostalosti biskupa Serdaheliho bol knižný prírastok početnejší, pozostával zo 152 zväzkov kníh. Na základe záznamov druhého katalógu bol v päťdesiatych rokoch 19. storočia knižničný fond rozšírený zhruba na 5 200 zväzkov. V tomto období zaznamenala knižnica väčší prírastok len v roku 1876, išlo o pozostalosť kanonika Tomáša Červeňa (1793 – 1876) v počte 384 kníh. Neskoršie údaje o rozsahu knižnice dokumentujú len uhorské štatistiky. V osemdesiatych rokoch 19. storočia na základe zistení Krajinského štatistického úradu sa fond rozšíril na 6 922 zväzkov.[10] Akvizíciu väčšieho rozsahu zaznamenala knižnica po smrti cirkevných predstaviteľov (František Berlica, Anton Majovský, Vincent Zeisl) a vo forme darov (František Staňák, Ivan Poliakovič).[11] Posledné údaje o stave fondu sú zo začiatku 20. storočia, keď predstavoval podľa zistení Krajinského dozorníctva múzeí a knižníc okolo 8 000 zväzkov. Tento nárast mohol byť spôsobený aj zjednotením viacerých cirkevných knižníc do spoločnej diecéznej knižnice. Stalo sa tak v roku 1909 spojením seminárnej, kapitulskej a krupinskej piaristickej knižnice s umiestnením v priestoroch predošlej seminárnej knižnice.[12]

8 MÉSZÁROSOVÁ, Klára: Katalóg jezuitskej knižnice v Banskej Bystrici z roku 1778. In: *Kniha 95 – 96. Zborník o problémoch a dejinách knižnej kultúry*. zost. M. Domová, Martin : Matica slovenská, 1997, s. 148 – 157.

9 Štátna vedecká knižnica v Banskej Bystrici, fond starých a vzácnych dokumentov, sign. MS Ba H 125: Index Primus Libros Bibliothecae Vener. Seminarii Neosoliensis in Scientiarum Ordines distributos exhibens, s. 9 (T2): 1852 E liberalitate Emerici Valij administri Hajnikiensis accessiti Bernolak Lexicon Exemplar unum in sex Tomis; s. 200: Post fata Rmi Dni Dominici Benedict Cath. Eccl. Theol. Lectoris et Canonici illati.

10 GYÖRGY, Aladár (ed.): *Hivatalos statisztikai közlemények: magyarsrszág köz- és magánkönytárai 1885-ben*, s. 276.

11 *Magyar minerva a magyarországi múzeumok és könyvtárak címkönyve. 5. kötet: 1912 – 1913*. Budapest : Athenaeum, 1915, s. 38.

12 *Magyar minerva a magyarországi múzeumok és könyvtárak címkönyve. 4. évfolyam: 1904 – 1911*. Budapest : Athenaeum, 1912, s. 65.

Seminárna knižnica bola umiestnená v jednej z miestností kňazského seminára (dnes budova Diecézneho centra Jána Pavla II. na Kapitulskej ulici). Nachádzala sa v maľovanej izbe, jej inventár tvorili skrine, v ktorých boli uložené knihy a podlhovastý stôl s dvoma stoličkami.

O knižnicu sa staral študijný prefekt a poverení študenti. V prvej polovici 19. storočia študijnými prefektmi boli: Peter Kern, Matej Kováč, Ján Tichý, Jozef Mozór, Ján Černák, Jozef Kozáček, Ján Országh, Štefan Záhorský, František Berlica a Jozef Munkay.[13] Viedli knižnicu v čase platnosti jej prvého knižničného katalógu a niektorí z nich v ňom pravdepodobne aj robili úpravy.

Za najdôležitejších spomedzi nich môžeme označiť prvých dvoch: Petra Kerna, ktorý vykonával tento úrad v rokoch 1807 – 1812 a 1819 – 1824, a jeho nástupcu Mateja Kováča, ktorý ním bol v rokoch 1812 – 1818. Kováč bol správcom knižnice aj v roku 1829, keď už ako kanonik a farár mesta Banská Bystrica predkladal katalóg knižnice biskupovi pri vizitácii. Jedného z nich je možné označiť aj za autora a zostavovateľa prvého katalógu knižnice.

Peter Kern (? – 1841) pochádzal z Kremnice. Po skončení teologických štúdií v roku 1805 sa vrátil do diecézy. Najprv pôsobil ako biskupský kaplán a protokolista u biskupa Serdaheliho. V novozriadenom seminári účinkoval od jeho vzniku v roku 1807 ako prefekt seminára, od roku 1808 bol i profesorom cirkevných dejín a dogmatiky a v roku 1812 sa stal jeho vicerektorom. Od roku 1824 pôsobil ako farár vo Veľkej Lovči, kde sa stal asesorom a neskôr dekanom dištriktu.[14]

Jeho nástupcom vo funkcii študijného prefekta a správcu knižnice sa stal Matej Kováč (1790 – 1848), ktorý pochádzal z Banskej Štiavnice. S pôsobením v seminári bol spojený už od roku 1812, patril medzi jeho prvých klerikov. V roku 1811 nastúpil ako aktuár do biskupského úradu k biskupovi Serdahelimu, neskôr sa stal jeho tajomníkom. V štruktúrach banskobystrického seminára zastával postupne niekoľko funkcií. Okrem funkcie študijného prefekta bol i profesor morálky a pastorálky, neskôr v rokoch 1843 – 1844 jeho rektorom. Okrem toho bol i banskobystrickým farárom od roku 1824 a riaditeľom miestneho gymnázia od roku 1832.[15]

Určení študenti sa v starostlivosti o knižnicu striedali pravdepodobne každý školský rok, pričom mali na starosti aj jej výpožičný denník. Menom sú známi len niektorí. V školskom roku 1838/39 ním bol študent 3. ročníka Ferdinand Bittera, v roku 1839 študent 4. ročníka Jozef Trost, v roku 1840 študent 4. ročníka Karol Pausch a v roku 1841 študent 3. ročníka teológie Štefan Schvandtner.[16]

Kňazský seminár sv. Karola Boromejského v Banskej Bystrici tak ako všetky diecézne semináre bol zrušený v roku 1950 na základe vládneho nariadenia č. 112/50. Jeho

13 *200 rokov Kňazského seminára v Banskej Bystrici: 1807 – 2007*. Banská Bystrica ; Badín : Kňazský seminár sv. Františka Xaverského, 2007, s. 64 – 65.

14 *200 rokov Kňazského seminára v Banskej Bystrici: 1807 – 2007*, s. 19 – 20.

15 *200 rokov Kňazského seminára v Banskej Bystrici: 1807 – 2007*, s. 17.

16 TATÁRIKOVÁ, Monika: Katalóg knižnice banskobystrického biskupského seminára sv. Karola Boromejského zo začiatku 19. storočia. In: *Kniha 2019. Zborník o problémoch a dejinách knižnej kultúry.* zost. D. Škulová, Martin : Slovenská národná knižnica, 2019, s. 206.

knižnica ako aj písomnosti seminára boli následne na vojenských nákladných autách[17] odvezené do Bratislavy. Podľa kultúrneho historika Jozefa Kuzmíka bola seminárna knižnica jednou z tých, ktoré postihli rozsiahle škody pri prevzatí štátnou mocou.[18] Jej novou správkyňou sa stala Matica slovenská, a to na základe darovacej zmluvy z roku 1958 so Slovenským náboženským fondom ako účelovým zariadením Slovenského úradu pre veci cirkevné. Fond knižnice vzhľadom na svoju rozsiahlosť, jednoznačné signatúry a vlastnícke záznamy zostal zhruba celistvý až na vyčlenenie rukopisných dokumentov a nepostihol ho osud menších knižníc, ktoré boli pomiešané do neidentifikovateľného súboru.[19] V roku 1968 bol skatalogizovaný na miestnom pracovisku Matice slovenskej. Jednotlivé knihy vtedy dostali novú signatúru MS Ba BB s počtom 6 552 jednotiek. Vytvorený bol aj lístkový autorský a zväzkový katalóg. Pôvodný katalóg knižnice ako rukopisný dokument dostal v tom čase signatúru MS Ba H 125 fol. V roku 1974 bola knižnica prevezená do Martina. V osemdesiatych rokoch 20. storočia bol jej fyzický stav až na niektoré jednotky pomerne dobrý aj napriek viacnásobnému premiestňovaniu, nešetrnému zaobchádzaniu s knihami a nevhodným podmienkam uskladnenia.[20] V roku 1998 bola v rámci reštitúcií vrátená pôvodnému majiteľovi Rímskokatolíckemu biskupskému úradu v Banskej Bystrici. V súčasnosti fond knižnice bývalého Kňazského seminára sv. Karola Boromejského spravuje na základe dohody s vlastníkom Štátna vedecká knižnica v Banskej Bystrici.

Rukopisný katalóg knižnice ako prameň výskumu

Vzhľadom na pohnutú históriu knižnice a jej fondu v minulosti má katalóg z čias jej vzniku veľký význam. Tento jedinečný rukopis je jedným z mála dochovaných prameňov, pretože väčšina zásadných písomností o chode seminára bola zničená v päťdesiatych rokoch 20. storočia. V prvom rade tak jeho dôležitosť spočíva v jednoznačnej dokumentácii a potvrdení existencie seminárnej knižnice v dejinách seminára.

Z hľadiska knihovníckej problematiky sa katalóg radí do oblasti knihovníckej dokumentácie. Bol určený na základnú evidenciu a požičiavanie kníh v seminárnej knižnici. Na rozdiel od súpisov majetkovej povahy, zaznamenávajúcich často len počty kníh, ktoré ich obsah evidovali len rámcovo. Knižničný katalóg s väčšou alebo menšou určitosťou identifikuje podľa obsahu alebo vnútorných znakov knihy alebo ich jednotlivé časti.

Katalóg tak predstavuje pozoruhodný a mnohými smermi výpovedný prameň. Možnosti jeho využitia sú ale determinované viacerými nadväzujúcimi prvkami. Patrí medzi ne spôsob, akým bol katalóg vytvorený, aké údaje obsahuje, dokonca i stupeň čitateľnosti zápisu a v neposlednom rade i úplnosť údajov v záznamoch. Prax pri tvorbe katalógov

17 *200 rokov Kňazského seminára v Banskej Bystrici: 1807 – 2007*, s. 89.

18 KUZMÍK, Jozef: Historické knižnice na Slovensku a ich postavenie v našej kultúrnej minulosti a prítomnosti. In: *Letopis pamätníka slovenskej literatúry*, 1968, s. 175 – 176.

19 AGNET, Ján: Historické knižné fondy na Slovensku: (genéza, vývoj a súčasný stav). In: *Teória a výskum knihovníctva a bibliografie. Výskumy č. 40: historické knižničné fondy na Slovensku*. Martin : Matica slovenská, 1988, s. 17 – 30.

20 SABOV, Peter: Historické knižničné fondy v správe SNK Matice slovenskej. In: *Teória a výskum knihovníctva a bibliografie. Výskumy č. 40: historické knižničné fondy na Slovensku*. Martin : Matica slovenská, 1988, s. 39.

bola totiž v minulosti veľmi kolísavá. Súviselo to nielen s akceptovanými pravidlami popisu, ktoré počas stáročí prakticky neexistovali, ale tvorcami katalógov mohli byť rôzne osoby rozličnej vzdelanostnej úrovne s rozličnými jazykovými schopnosťami.

Súpis kníh knižnice je základným a najdôležitejším pramenným materiálom na výskum širších súvislostí dejín knižnej kultúry, ale aj jej čiastkových disciplín. Medzi prvoradé patria predovšetkým dejiny samotnej knižnice. Jednak prostredníctvom analýzy jej fondu a čitateľov, jednak prostredníctvom skúmania regionálnych prvkov dejín knihovníctva. Pri stanovených hľadiskách je potrebné prihliadať na špecifickú kategóriu, akou bola seminárna knižnica, ktorá sa radila v klasifikácii knižníc zároveň medzi školské i cirkevné formy knižníc.

Charakter knižnice je vhodné vziať do úvahy hlavne pri posudzovaní jej fondu, pretože najmä tu sa odzrkadľuje príslušnosť knižnice k inštitúcii. Zo skladby seminárneho fondu možno získať primárne dvojaké údaje. Jednou z možností je zhodnotenie z aspektu jej obsahu. V tomto prípade je výskum zameraný na samotné knihy, pričom si kladieme otázky ohľadom ich zamerania, počtu vydaní konkrétnej publikácie, ich popularity v určitom období, resp. zisťovania informácií o ich autoroch. Možno z nich získať údaje aj o tom, čo v určitom čase ponúkal domáci knižný trh a aké knihy sa dovážali zo zahraničia. Ďalšou možnosťou je prostredníctvom kníh usudzovať o záujme a potrebách čitateľov knižnice o určité knihy alebo skupiny kníh, typ literatúry, v širšom chápaní aj o jej využiteľnosti a čitateľských návykoch konkrétnej skupiny dobových čitateľov, akými boli seminaristi a ich vyučujúci. Takto získané údaje informujú o úrovni knižničnej zbierky, ktorej skladba odráža možnosti jeho dopĺňania, resp. podriadenosť zameraniu seminárnej knižnice.

Prameň môže byť aj objektom výskumu knihovníckej práce v minulosti, pretože je vzácnym zdrojom údajov o spôsobe práce v knižnici a jej knihovníkoch. Jeho rozbor z tohto pohľadu prináša nové poznatky o použitom type triedenia, organizácii fondu, o prírastkoch a o výpožičkách, ale i o štýle vedenia záznamov a katalogizačnej praxi v priebehu viacerých desaťročí. Do tejto kategórii informácií môžeme zaradiť aj paleografický výskum rukopisov jednotlivých knihovníkov. Keďže ide o text domácej, slovacikálnej proveniencie, hodnotiť môžeme vzhľad písma a celkový obraz písma, ktorý dokumentuje vývoj písma v našom kultúrnom prostredí 19. storočia.

Seminárna knižnica sa vo svojom pôvodnom zameraní vyvíjala a využívala v minulosti, zo strany bádateľov v prítomnosti sa ťažisko jej významu prenáša na historicky cenné fondy, ktoré sú základňou výskumu dejín knižnej kultúry. Prínos pôvodného katalógu seminárnej knižnice spočíva teda v svedectve dokumentácie jej pôvodného stavu, ktorý existoval pri zakladaní knižnice a v prvých rokoch jej pôsobenia. Jeho prebádanie môže napomôcť aj vyriešenie niektorých nejednoznačných správ z tohto obdobia. Získané informácie z oblasti profilu knižničného fondu, čitateľov a knihovníckej praxe otvárajú aj nové možnosti ďalšieho výskumu, ktorý bude vytvárať kooperácia pôvodných údajov v porovnaní s existujúcim zachovaným fondom a jeho provenienčnými záznamami. Dôkladný rozbor je zdĺhavá a náročná úloha, viaceré z načrtnutých výskumných úloh kladú dôraz na presné štatistické vyhodnotenie katalógu, k čomu výraznou mierou môže dopomôcť automatická transkripcia rukopisu a hlavne ďalšie možnosti platformy *Transkribus*.

ELENCHUS LIBRORUM

Vonkajší vzhľad rukopisného prameňa predstavuje formát B4 s rozmermi 38 x 22 x 5cm. Je viazaný v pôvodnej kombinovanej väzbe. Lepenkové knižné dosky presahujúce knižný blok sú potiahnuté ozdobným farebným papierom, pričom jeho rohy a chrbát sú pokryté aj hnedou kožou, ktorá je zdobená jednoduchou i dvojradovou linkovou slepotlačou nachádzajúcou sa na okraji presahu s papierovou časťou väzby. Na prednej knižnej doske je prilepená ozdobne vystrihnutá nálepka z bieleho papiera, ktorá je umiestnená v jej hornej polovici. Obsahuje latinský štrnásťriadkový rukopisný text, ktorý sa začína slovami Elenchus Librorum. Pravdepodobne popisovala obsah dokumentu. Zvyšok nápisu je v súčasnosti vplyvom potečenia vodou ťažko čitateľný. V hornej časti zadnej knižnej dosky je umiestnená knižničná nálepka so signatúrou MS Ba H 125 fol. Na chrbte väzby sa nachádzajú štyri dobre zachované väzy, po bokoch ozdobené linkovou slepotlačou, podobne ako na oboch knižných doskách. Forma väzby prostredníctvom viditeľných vypuklých väzov odkazuje na nemecký typ väzby. Väzba a knižný blok sú zošité konopnými niťami, ktoré sú rozpoznateľné pre mierne poškodenie väzby. Oriezka je husto postriekaná červenou farbou. Formálnu stránku dokumentu dopĺňajú aj konopné šnúrky a malé papierové štítky, ktoré predstavujú prepojenie fyzickej stránky prameňa s jeho vnútorným obsahom. Konopné šnúrky sú pripevnené na prvých listoch dvoch samostatných častí dokumentu, prvého knižničného súpisu a výpožičného denníka. Následne rovnako na vonkajších okrajoch listov sú nalepené papierové štítky, ktoré predstavovali začiatok jednotlivých skupín druhého knižničného súpisu. Obsahujú v abecednom poradí rukopisné kapitálky písmen A až J, pričom písmeno K pre poslednú tematickú skupinu bolo v minulosti odtrhnuté. Uvedené prvky slúžili na uľahčenie orientácie a rýchlejšie vyhľadávanie v dokumente. Do veľkej miery ošúchaný vonkajší vzhľad a zodratý materiál s ryhami a škvrnami po tuši svedčí o pravidelnom a dlhodobom používaní dokumentu.

Knižný blok pozostáva zo súboru papierových knižných zložiek, ktoré sú tvorené listami kvalitného ručného papiera. Papier na základe použitých priesvitiek pochádzal z papierne v Kamenci pod Vtáčnikom (okres Prievidza).[21] V dokumente sa striedajú dva varianty identifikačného znaku tejto papierne. Prvým je štylizovaný erb rodiny Kostolániovcov, ktorý okrem bežných motívov často používali majstri tejto papierne. Zobrazuje v štíte na pažiti pred košatým stromom doprava kráčajúceho medveďa, v horných rokoch štítu sú privrátený polmesiac a hviezda. Klenot tvorí doprava obrátený medveď. Druhý variant priesvitky tvoria kapitálky písmen A a R. Postupnosť oboch variantov sa opakovala v poradí jeden erb, jedna dvojica písmen alebo dva erby, dve dvojice písmen. Na základe použitia konkrétneho typu priesvitiek bolo možné určiť presnejšie časové obdobie vzniku papiera. Je ním obdobie medzi rokmi 1796 až 1802, keď bol papierenským majstrom kameneckej papierne Adalbert Rosenburger.

Prameň obsahuje spolu 512 strán, pričom sa medzi nimi nachádzajú aj viaceré vakanty. Prázdne strany, ktoré mali slúžiť na budúce prírastky, sa vyskytujú jednak medzi jednotlivými skupinami, jednak medzi samostatnými časťami. Očíslovanie prameňa arabskými číslicami nebolo pôvodne úplné. Najprv bol po stranu 273 čiernym tušom označený rozsah strán prvého katalógu. Následne sa pokračovalo ceruzkou pravdepodobne až v priebehu 20. storočia pri katalogizácii dokumentu.

21 DECKER, Viliam: *Dejiny ručnej výroby papiera na Slovensku*. Martin : Matica slovenská, 1982, s. 50, 133.

Vnútorná štruktúra dokumentu pozostáva z viacerých samostatných častí, ktoré boli vytvorené postupne v chronologickej následnosti. Medzi hlavné patria dva knižničné katalógy, výpožičný denník a akvizičné zoznamy. Odlišujú sa vizuálne, obsahovo i obdobím svojej pôsobnosti.

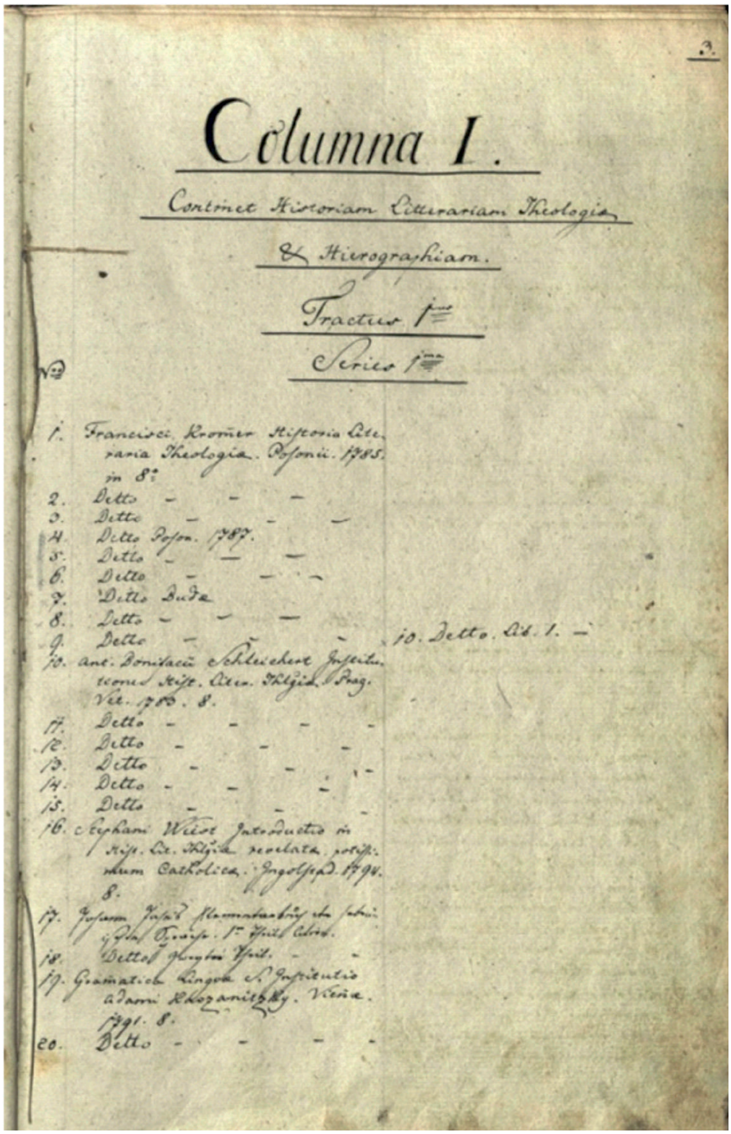
INDEX PRIMUS A JEHO PÍSMO

Na titulnom liste dokumentu je uvedený názov *Index primus libros bibliothecae vener. seminarii Neosoliensis in scientiarum ordines distributos exhibens*, ktorý reprezentuje na 137 listoch prvý nedatovaný katalóg seminárnej knižnice. Základné členenie katalógu predstavuje desať tematických tried, z ktorých každá je označená rímskou číslicou a jej textovým popisom:

Columna I. Continet Historiam Litterariam Theologiae & Hierographiam
Columna II. Continet Patrologiam, & Theologiam Dogmaticam, eamque a) Scholasticam & b) Catechet.
Columna III. Continet Theologiam Dogmaticam eamq. a) Scholasticam, b) Catecheticam & Theol. Polemicam
Columna IV. Continet Theologia Moralem & Asceticam
Columna V. Continet Theologiam Pastoralem, & Homileticam
Columna VI. Continet Theologiam Homileticam
Columna VII. Continet Theologiam Lyturgicam
Columna VIII. Continet Theologiam Lyturgicam & Historiam tam Sacram quam profanam
Columna IX. Continet Historiam, Jurisprudentiam, Phylosophiam, Phylologiam &c Varii denique argumenti
Columna X. Continet Jurisprudentiam, Phylosophiam, Phylologiam, Pedagogiam &c varii denique argumenti

Členenie katalógu kopírovalo uloženie kníh v knižnici. Hlavné tematické triedy sa začínajú vždy na samostatnej strane, pričom sa ďalej členia na dve podskupiny, tractus (spravidla 5 – 7) a series (spravidla 1 – 2), ktoré nesú už len číselné označenie arabskou číslicou bez bližšieho popisu. Strany v katalógu sú dvoma vertikálnymi čiarami znázornenými ceruzkou rozdelené na tri stĺpce. Prvý, užší slúžil len na zápis poradového čísla záznamu, pričom číslovanie bolo samostatné pre každú podskupinu. Zvyšná časť strany bola rozčlenená na dva rovnaké stĺpce, ktoré zachytávali záznamy o knihách. Vo všeobecnosti platí, že jednému záznamu v katalógu prislúcha jedna knižničná jednotka. Katalóg tak obsahuje zhruba 3 600 záznamov o knihách. Ľavý stĺpec sa vyplnil hneď pri zostavovaní katalógu. Jeho záznamy boli značené v jazyku dokumentu, pričom zachytávajú celkom podrobné informácie o knihe ako takej. Obsahujú informácie o autorovi, názve, vydavateľských údajoch a formáte dokumentu. Autor diela je uvedený zväčša v genitívnej podobe mena. Názov diela je často krátený, prípadne upravený. V prípade viaczväzkových diel sa doplnil aj údaj o konkrétnej časti zväzku, v prípade konvolútov aj údaje o jednotlivých príväzkoch. Stávalo sa, zrejme vplyvom nedbalosti knihovníka, že zápis mena autora alebo názov diela bol v porovnaní s konkrétnou knihou nepresný a skomolený. Vysvetlením môže byť aj hypotéza, že na tvorbe katalógu pracovali dvaja knihovníci, pričom jeden diktoval údaje z knihy a druhý robil jej zápis.

Vydavateľské údaje zahrňujú miesto vydania a rok. V zápisoch sa z úsporných dôvodov využívali aj skratky slov, skracovali sa hlavne slová v názve diela a miesto vydania. Posledný stĺpec bol vyhradený pre informácie o zmenách v katalógu v priebehu jeho trvania. Záznam o knihe bol v tomto prípade doplnený o vysvetlenie zápisu, poprípade časové upresnenie a zriedka i meno predošlého vlastníka. Najčastejšie sa sem značila akvizícia kníh do fondu, či už išlo o kúpu, dar alebo pozostalosť, ale tiež straty a časté boli aj výmeny medzi jednotlivými tematickými skupinami alebo podskupinami. Datované doplnky pochádzali z rokov 1813 – 1816, 1848, 1850 – 1852.



3.

Columna I.

Continet Historiam Litterariam Theologiae & Hierographiam.

Tractus 1us

Series 1ma

No

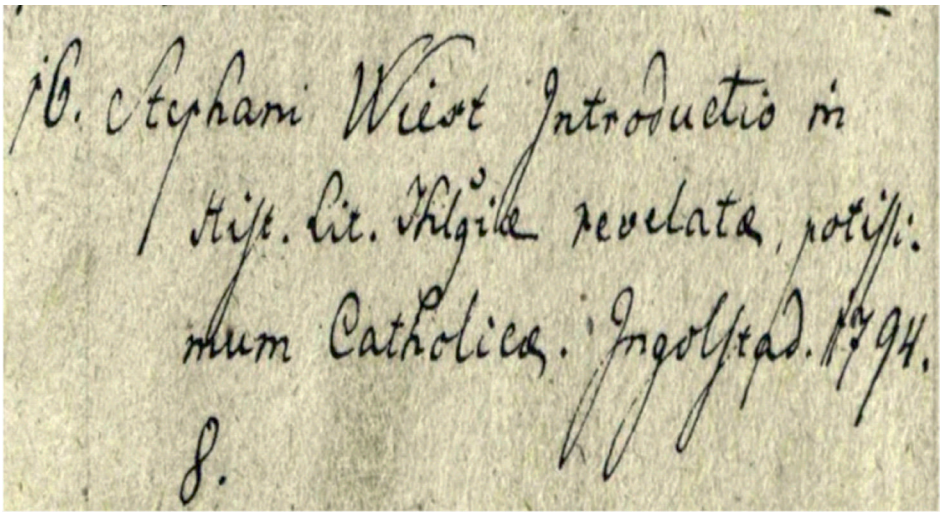
1. Francisci Kromer Historia Litteraria Theologiae. Posonii. 1785. in 8o
2. Detto
3. Detto
4. Detto Poson. 1787.
5. Detto
6. Detto
7. Detto Buda
8. Detto
9. Detto — 10. Detto. Lib. 1.
10. Ant. Bonifacii Schleichert Institutiones Hist. Liter. Theologiae. Prag. Vet. 1780. 8.
11. Detto
12. Detto
13. Detto
14. Detto
15. Detto
16. Stephani Weiss Introductio in Hist. Lit. Theologiae revelatae potissimum Catholicae. Ingolstad. 1794. 8.
17. [illegible]
18. Detto [illegible]
19. Gramatica Lingvae S. Institutio Adami Kkozanitzky. Viena. 1791. 8.
20. Detto

Obrázok 48 Prvá strana katalógu. Zdroj: Štátna vedecká knižnica v Banskej Bystrici.

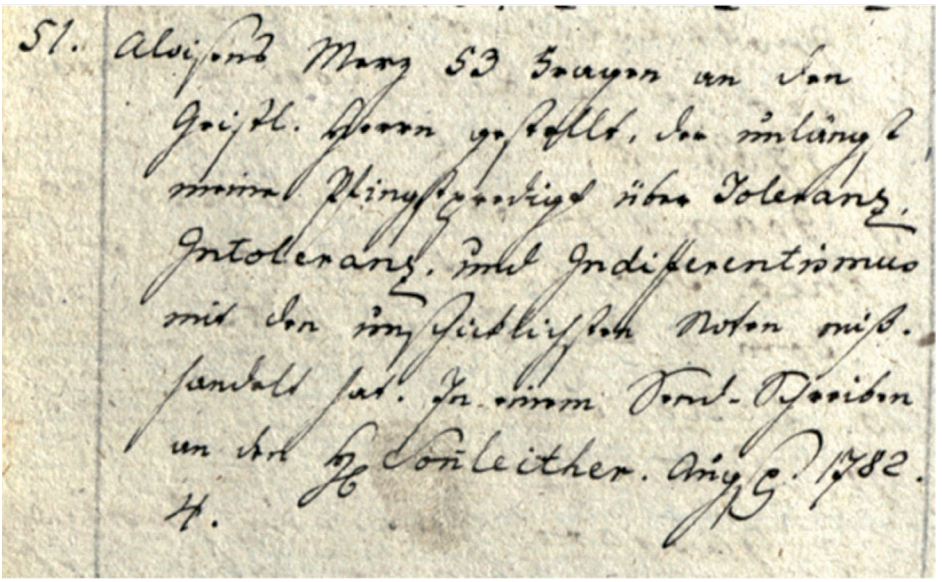
Knihovníci pri tvorbe katalógu uplatnili aj ďalšie zásady knihovníckej práce, ktoré sa týkali využívania špeciálnych knihovníckych výrazov, skratiek a symbolov. Výraz „detto“ bol použitý na zápis duplikátov alebo pri zhodných častiach názvu viaczväzkových diel. Výraz „item“ označoval rozpis príväzkových diel. Výraz „sine tit.“ sa uplatňoval pri chýbajúcom údaji. Ďalšie knihovnícke symboly predstavovali napríklad poznámku ku knihe, akvizičný doplnok alebo výmenu medzi podskupinami. Na dôležitú informáciu upozorňovala aj skratka NB pochádzajúca z latinského výrazu „nota bene“.

Prvý katalóg seminárnej knižnice sa používal zhruba päťdesiat rokov. Vyplýva to jednak z časových údajov jeho doplnkov, jednak z datovania druhého katalógu. Za tento čas sa v jeho správe vystriedalo niekoľko knihovníkov. Paleografická analýza písma rozlíšila viacero rukopisov, ktoré sa však striedali len v zápise doplnkov (identifikované boli zatiaľ minimálne štyri). Pôvodný tvorca katalógu ho mal na starosti do roku 1813, pre záznamy od roku 1815 už bola použitá ďalšia pisárska ruka. Vo všeobecnosti ide v priebehu celého katalógu o úhľadne napísaný a dobre čitateľný text zapísaný čiernym dubienkovým atramentom. Toto pravidlo už neplatí pre texty doplnkov, ktoré sú v niektorých prípadoch horšie čitateľné.

Základ katalógu bol vytvorený začiatkom 19. storočia, keď v našej kultúrnej oblasti súbežne popri sebe fungovali viaceré typy písma pre konkrétne jazyky. Jazyku zápisov tak zodpovedajú zaužívané aj vizuálne odlišné typy písma so špecifickými rysmi, teda dominantnejšie zastúpené humanistické kurzívne písmo pre latinčinu (a iné jazyky) a kurent (novogotická kurzíva) pre nemčinu. Ojedinele sa v texte vyskytujú aj novogotické polokurzívne znaky pre češtinu a grécka alfabeta. Dva základné typy písma sa nepravidelne striedajú v celom dokumente, v niektorých prípadoch aj samotný záznam pozostával z oboch typov, ale ide o jednu pisársku ruku. To platí len pre ľavý stĺpec, pravý s doplnkami bol zhotovený viacerými pisármi.



Obrázok 49 Ukážka humanistického kurzívneho písma. Zdroj: Štátna vedecká knižnica v Banskej Bystrici.



Obrázok 50 Ukážka novogotickej kurzívy. Zdroj: Štátna vedecká knižnica v Banskej Bystrici.

Za prvým katalógom na stranách 277 – 284 sa nachádza prvý rozsiahlejší nekatalógový súpis kníh, ktorý sa začína slovami: *Elenchus librorum qui in Bibliothecis episcopalibus neosoliensi ac S. Crusensi in duplici exemplari reperti.* Pozostáva zo stopäťdesiatdva kníh, ktoré boli na základe testamentu biskupa Serdaheliho včlenené do seminárnej knižnice. Skladá sa z dvoch častí, prvú tvorili duplikáty kníh z biskupskej knižnice v Žiari nad Hronom, druhú tvorili duplikáty z knižnice v Banskej Bystrici. Jednoduchý súpis bol vytvorený po roku 1813 už inou pisárskou rukou ako prvý katalóg. Záznamy o knihách obsahujú len názov knihy, meno autora a číslo časti. Prevažne ide o latinsky písané diela s náboženskou tematikou, ale spoločným znakom ďalších je uhorská tematika rôznych odvetví.

Dôležitou súčasťou prameňa je výpožičný denník knižnice, ktorý zahŕňa strany 287 – 352. *Elenchus librorum extradatorum* obsahuje približne 1500 záznamov o vypožičaných knihách z rokov 1837 – 1876, čiže sa používal počas platnosti oboch knižničných katalógov. Pozostával však z dvoch častí, keďže počas pôsobnosti druhého katalógu sa zmenilo usporiadanie knižnice. Správna metodika zápisu výpožičiek pre meniacich sa pomocných knihovníkov je zaznačená v jeho úvode, pričom obsahuje aj mená vtedajších správcov. Za študentov ním bol Ferdinand Bittera, študent 3. ročníka teológie, a za profesorskú časť prefekt štúdií František Lopušný (1813 – 1874). Denník fungoval v podobe tabuľky, do ktorej sa značili údaje o mene požičiavateľa, o autorovi a názve vypožičaného diela, o mieste uloženia a dátume výpožičky. Preškrtnutý zápis znamenal, že kniha bola vrátená do knižnice.

Druhý katalóg knižnice, ktorý je rozpísaný na stranách 355 – 499, bol zostavený v roku 1857 pod názvom *Index novus libros bibliothecae venerab. seminarii Neosoliensis exhibens 1857*. Z predošlého katalógu bolo ponechané triedenie do desiatich

tematických skupín, ktoré sú v tomto prípade označené majuskulami A až K:

Columna A. Diversi argumenti Opera

Columna B. Opera Biblica

Columna C. Opera prophana diversi argumenti

Columna D. Vita Sanctorum, Opera Concionatoria nec non Vitae Jesu Christi

Columna E. Opera Concionatoria

Columna F. Opera Ascetica et Moralia

Columna G. Opera juridica

Columna H. Pastorales, Opera liturgica, ac juridica

Columna I. Opera Dogmatica

Columna K. Opera historica et Sanctorum Patrum

Rekatalogizácia knižnice priniesla viacero zmien v úprave a vedení katalógu. Strany boli troma vertikálnymi čiarami rozdelené do troch stĺpcov, pričom oba krajné boli užšie, najviac priestoru bolo ponechaného strednému stĺpcu. Prvý slúžil na zápis poradového čísla, druhý na zápis o knihe, tretí sa ojedinele využíval na značenie poznámok formou knihovníckych skratiek (napríklad krúžky, pomlčky, krížiky). Členenie v podskupinách ostalo len vo forme tractus. Odlišná je aj podoba zápisu záznamov, ktorá uvádza menej informácií a je neprehľadná. Doplnky a zmeny sa evidovali len na konci každej podskupiny a zväčša bez uvedenia dátumu alebo mena predošlého majiteľa. Zaznamenané zmeny sú z rokov 1860 – 1869 a 1874 – 1875. Počet kníh v knižnici sa zvýšil v období druhého katalógu na približne 5 200 zväzkov.

V závere prameňa na stranách 503 – 509 sú umiestnené viaceré nesúrodé informácie o chode knižnice datované do sedemdesiatych rokov 19. storočia. Rozlíšiť sa dajú niektoré súpisy prírastkov (biskup Arnold Ipolyi-Stummer, rektor seminára František Berlica), spomedzi nich vyniká zoznam 384 kníh z pozostalosti významného národnokultúrneho pracovníka a banskobystrického cirkevného hodnostára Tomáša Červeňa, datovaný do roku 1876.

Súčasným vlastníkom rukopisného prameňa *Elenchus librorum* je od roku 2000 Štátna vedecká knižnica v Banskej Bystrici, keď ho darom získala od predošlého vlastníka Rímskokatolíckeho biskupstva v Banskej Bystrici. Dokument je spracovaný v online katalógu knižnice pod signatúrou O 3792. Bádateľom je prístupný prezenčne za splnenia podmienok študovne starých a vzácnych dokumentov.

TVORBA MODELU V PLATFORME *TRANSKRIBUS*

Aplikácia nástrojov platformy *Transkribus* vyžadovala previesť rukopisný prameň *Elenchus librorum* do digitálneho grafického súboru.

Proces jeho digitalizácie prebiehal v priestoroch Štátnej vedeckej knižnice v Banskej Bystrici a riadil sa postupmi internej smernice upravujúcej proces digitalizácie knižničného fondu v Štátnej vedeckej knižnici v Banskej Bystrici. Digitálnu kópiu

vyhotovila pracovníčka oddelenia doplňovania a spracovania fondov. Originálny dokument bol nasnímaný na manuálnom planetárnom skeneri Bookeye 4, ktorý disponuje knižnou kolískou tvaru V, chrániacou knižnú väzbu v procese digitalizácie pred poškodením. Prvotné skeny jednotlivých strán dokumentu boli následne upravené v tzv. post-processingu, ktorý zabezpečuje s využitím softvérových programov vzhľad jednotlivých skenov. K uskutočneným úkonom patrí napríklad oprava orientácie, natočenie a orezanie strán, určenie okrajov strán a kontrola zoradenia strán. Na zachovanie autentickosti materiálu sa v tomto štádiu nastavuje aj farebné pozadie, respektíve pozadie v stupňoch šedej a nie bitonálne.

Digitalizát, ktorý tvorí kompletný rozsah vyššie špecifikovaného dokumentu, predstavuje celkom 260 snímok naskenovaných zdrojových obrázkov v rozlíšení 600 dpi vo formáte TIFF. Stanovené technologické parametre dopĺňa farebná hĺbka color24-bit.

Z celkového rozsahu digitalizátu bolo využívaných v tomto štádiu výskumu len 122 snímok prvého katalógu, pričom jedna snímka predstavuje jednu dvojstranu originálneho dokumentu. Do tohto počtu nie je zarátaných 18 odstránených snímok prázdnych dvojstrán.

Následnú prácu so snímkami prebiehajúcu v prostredí *Transkribus* ovplyvňujú špecifiká vybraného dokumentu. Prejavujú sa jednak v prípravných krokoch tvorby modelu, akou je segmentácia a vlastný prepis, jednak zasahujú do formovania samotného modelu.

Prvým znakom je formálny vzhľad dokumentu, ktorý obsahuje ručne predkreslené čiary stĺpcov. Hoci v tomto prípade nejde o klasickú tabuľku s pevne stanovenými blokmi, obsahuje aj nepravidelne sa vyskytujúce pohyblivé záhlavia. Jej charakter má ale zásadný vplyv pri rozčleňovaní jednotlivých snímok katalógu do textových polí a možnosť vytvoriť tabuľky vo fáze segmentácie má pozitívny účinok na tvorbu trénovacích údajov. Vo výsledku tak každej snímke zodpovedajú dve tabuľky rozdelené do troch stĺpcov, pričom každý záznam o knihe je v samostatnej bunke. Táto schéma je priebežne narúšaná záhlavím jednotlivých skupín a podskupín.

Existencia tabuliek ovplyvnila aj automatickú segmentáciu a vyžiadala si dodatočnú manuálnu segmentáciu a dôslednú kontrolu a korekciu. Nedostatky boli zrejme spôsobené aj tým, že samotný rukopis je bez pomocných lineárnych čiar a rozostupy medzi riadkami sú nerovnomerné. V takomto prípade sa často stávalo, že základné čiary viacerých slov „vychádzali“ z textových oblastí. Opravu si vyžadovali najčastejšie dve skupiny chýb. Prvou bolo neakceptovanie stĺpcov, čo malo za následok manuálne rozdeľovanie jednotlivých celostránkových riadkov. V jej nadväznosti bola nesprávna následnosť riadkov v stĺpcoch, čo ovplyvnilo správne poradie čítania buniek. Manuálnu korekciu potreboval aj text záhlaví, pri ktorom bolo nevyhnutné názvy jednotlivých skupín a podskupín spájať do spoločných buniek. Spájanie viacerých buniek som využila aj pri stĺpci doplnkov, keďže niektoré zápisy nebolo možné rozdeliť do samostatných riadkov aj z toho dôvodu, že záznamy neboli zväčša číslované. V stĺpci doplnkov preto jednej bunke zodpovedá spravidla celý záznam doplnku týkajúci sa jednej udalosti bez ohľadu na to, koľko obsahuje kníh.

Ďalším znakom je nesúlad pisárskych rúk v ľavom a pravom stĺpci. Tento fakt ovplyvnil následný prípravný krok tvorby modelu, ktorý predstavuje manuálny presný prepis časti textu rukopisu. Na tréning modelu som vybrala len základné záznamy v ľavom stĺpci, ktoré sú značené jednou pisárskou rukou pôvodného autora katalógu. Vzhľadom na to, že doplnkové záznamy v pravom stĺpci sú vytvorené viacerými pisárskymi rukami, som ich z tréningu vylúčila, a to ich označením metadátovým výrazom „unclear". Dôvodom ich vynechania z procesu tvorby modelu bola aj výrazná odlišnosť niektorých rukopisov oproti základnému z ľavého stĺpca a malý počet znakov, ktorý by nezabezpečil ich rozpoznanie v procese výuky modelu. Z tréningu modelu som prostredníctvom metadátového vyjadrenia vylúčila aj nejasné výrazy.

Vlastný prepis textu som uskutočnila najprv na deviatich snímkach prvej tematickej skupiny katalógu (Continet Historiam Litterariam Theologiae & Hierographiam). Z tohto počtu predstavoval cvičný súbor osem snímok s 2 876 slovami. Následne pri tréningu bol počet znížený na 2 702 slov odpočítaním riadkov, v ktorých sa nachádzali slová označené výrazom „unclear". Overovací súbor reprezentoval jeden obrázok so 414 slovami. V texte mala dominantné zastúpenie latinčina v humanistickom kurzívnom písme, hoci ojedinele sa aj tu vyskytol nemecký kurent.

V druhom kroku som uskutočnila výber dvanástich snímok takmer zo všetkých tematických skupín katalógu (mimo skupiny č. 7 a 8), v ktorých boli vo väčšej miere zastúpené záznamy v nemčine. Nepomer oboch výberov je zapríčinený menšou dostupnosťou nemeckého kurentu v záznamoch oproti humanistickému písmu. Pomer cvičného súboru s jedenástimi snímkami k overovaciemu súboru s jednou snímkou predstavoval v tomto prípade 4041 slov k 342 slovám.

Adekvátne využitie možností automatickej transkripcie platformy *Transkribus* si vyžaduje stanoviť vhodné pracovné postupy. Zásadný vplyv na tvorbu modelu rukopisného katalógu má jeho rozsah a použité typy písma.

Trénovanie funkčného modelu rozpoznávania rukopisu v platforme *Transkribus* sa odporúča so vzorkou 5 000 – 15 000 slov prepísaného materiálu. Vzhľadom na menší rozsah rukopisného katalógu by bolo potrebné na splnenie tejto požiadavky manuálne prepísať zhruba polovicu naskenovaných snímok, čo by bolo neefektívne. Preto som sa sústredila na možnosť využitia tzv. malého modelu vytvoreného na menšej tréningovej vzorke, ako navrhujú oficiálne príručky. V tomto prípade je ale potrebné nízke množstvo požadovaných tréningových dát nahradiť využitím vhodného, verejne dostupného základného modelu. Následne sa informácie, ktoré obsahuje základný model, integrujú do nového modelu, čo môže zlepšiť jeho výsledky. Dôležitou podmienkou prínosu základného modelu je podobnosť jeho textu s textom, ktorý má následne rozpoznať.

V neposlednom rade boli jednotlivé kroky tréningu konkrétneho modelu ovplyvnené najmä dvoma použitými odlišnými typmi písma. Experiment som z tohto dôvodu rozdelila na dve fázy. V úvodnej fáze som vytvorila samostatné modely pre oba dominantné typy písma, latinku a kurent. V rámci jednotlivých pokusov som oba modely skúšala v dvoch verziách. Jednu verziu som vytvorila len na základe vlastného prepisu textu a jednu verziu za použitia základného modelu.

V prípade latinky dosiahla chybovosť verzie s vlastným prepisom úroveň 4,44 %. Pre druhý variant som použila základný model s názvom „NeoLatin_Ravenstein_1643-1772“, ktorý bol vyvinutý pre jezuitské písomnosti z holandského mesta Ravenstein z obdobia rokov 1643 – 1772.[22] Samotný základný model bol pôvodne trénovaný na 64 435 slovách s chybovosťou overovacieho setu na úrovni 3,58 %. Vlastná verzia latinky sa s jeho pomocou zlepšila najprv na úroveň chybovosti 3,75 %, po revidovaní prepisu na úroveň 3,05 %. V porovnaní oboch verzií modelu to predstavovalo zdokonalenie o 1,39 %.

Pri trénovaní modelu nemeckého kurentu mala vlastná verzia modelu úroveň chybovosti 4,37 %. V prípade modelu pre kurent som urobila pokus s dvoma základnými modelmi, „Kurrent_1515_ENHG_v3“ a „StAZH_RRB_German_Kurrent_XIX“. V platforme *Transkribus* je síce k dispozícii viacero modelov nemeckého jazyka, náhľad textov ale neposkytujú všetky, a preto je komplikované zvoliť vhodný základný model. Vyberala som z takých modelov, ktoré mali percento chybovosti nižšie ako môj vlastný model. Lepšie hodnoty s úrovňou chybovosti 3,39 % zaznamenal novovytvorený model za pomoci základného modelu s názvom „StAZH_RRB_German_Kurrent_XIX“. Tento model bol pôvodne trénovaný na švajčiarskych rukopisných zápisniciach z 19. storočia z prostredia Zürichu na vzorke až 26 026 908 slov. Chybovosť overovacieho súboru predstavovala hodnotu 1,73 %. Rozdiel úrovne chybovosti v porovnaní dvoch verzií môjho modelu kurentu znamenal 0,98 %.

Zo vzájomného porovnania dvoch nových modelov pre jednotlivé typy písma dosiahol lepšie výsledky model pre latinský jazyk o 0,34 %, ale oba modely aj napriek menšiemu počtu slov zaznamenali dobré výsledky vďaka použitiu vhodných základných modelov. Vytvorené modely sa stali podkladom pre druhú fázu, v ktorej som spojila všetky prepísané snímky z predošlých modelov. Porovnanie oboch výsledkov zhodnotí úspešnosť pokusu.

Do novej fázy výskumu tak bolo na začiatku zapojených 19 snímok cvičného súboru a v overovacom súbore boli umiestnené dve snímky, jedna pre latinčinu a druhá pre nemčinu (snímka pre nemčinu obsahuje približne 2/3 záznamov v nemčine). Dokopy tak bolo použitých zhruba 17 % z celkového počtu snímok prvého katalógu. Počet prepísaných slov sa zvýšil na 6 755 slov v cvičnom súbore a 2 064 slov v overovacom súbore. Spoločný model oboch typov písma som vyhotovila tiež v dvoch verziách ako v predošlej fáze výskumu. Chybovosť bola opäť vyššia pri vlastnom prepise, a to na úrovni 3,68 %. V druhej verzii som si za základný model opäť vybrala už raz použitý model pre nemčinu „StAZH_RRB_German_Kurrent_XIX“. Uprednostnila som tento výber pre rovnovážnejšie zastúpenie oboch typov písma v spoločnom modeli. V tejto verzii modelu dosiahla výsledná chybovosť uspokojivú úroveň 3,18 %, čo predstavovalo zlepšenie o 0,50 %. Pri podrobnejšej analýze oboch snímok overovacieho súboru vykazovala nižšiu chybovosť snímka s latinským textom (2,98 %) oproti nemeckému kurentu (3,03 %), ale predstavovalo to len zanedbateľný rozdiel.

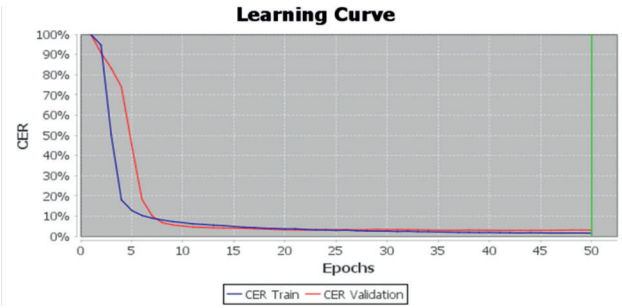
22 Model „NeoLatin_Ravenstein_1643-1772“ je aplikovaný aj v modeloch uvádzaných v kapitolách, 3, 4 a 6.

Tabuľka 9 Prehľad vytvorených modelov (Elenchus librorum).

Model č.	ID	Jazyk	Počet slov	Počet riadkov	Základný model	ID	CER %	
							Cvičný súbor	Overovací súbor
1	43550	Latinský	2702	876	NeoLatin_Ravenstein_1643-1772	22408	0,36	3,75
2	43563	Latinský	2702	876			0,34	4,44
3	43604	Latinský	2714	880	NeoLatin_Ravenstein_1643-1772	22408	0,45	3,05
4	44292	Nemecký	4041	1184			0,83	4,37
5	44400	Nemecký	4041	1184	Kurrent_1515_ENHG_v3	40510	0,77	3,54
6	44402	Nemecký	4041	1184	StAZH_RRB_German_Kurrent_XIX	30919	0,69	3,39
7	44901	Latinský+Nemecký	6755	2064			1,2	4,16
8	44907	Latinský+Nemecký	6755	2064			1,24	3,68
9	44919	Latinský+Nemecký	6755	2064	StAZH_RRB_German_Kurrent_XIX	30919	1,49	3,18
10	44979	Latinský+Nemecký	841	252	Skriptor_library_catalog	44919	0,19	5
11	44980	Latinský+Nemecký	7596	2316			1,49	3,26
12	45004	Latinský+Nemecký	7596	2316	StAZH_RRB_German_Kurrent_XIX	30919	1,5	3,08

Výsledný spoločný model s identifikačným číslom 44919 som overila už za pomoci automatického prepisu na ďalších štyroch snímkach dokumentu, ktoré pochádzali zo štyroch rôznych tematických skupín katalógu so zastúpením oboch typov písma. Chybovosť na vybraných snímkach tvorila úroveň v rozmedzí 2,72 % – 3,38 %. Výsledný automatický prepis je uspokojivý, prepis textu záznamov v oboch jazykoch čitateľný.

Na základe tohto pokusu som mala ešte snahu zdokonaliť existujúci model 44919 s použitím opravených snímok po predošlej automatickej transkripcii. Na vytvorenie pokusných modelov som využila tri postupy. V prvom prípade súbor tréningových údajov obsahoval len šesť snímok so základným modelom 44919, po tri pre cvičný aj overovací súbor. Výsledok presnosti modelu bol na úrovni 5 %. Pri druhom pokuse som aplikovala všetkých 25 snímok, z toho 22 predstavovalo cvičný súbor so 7 596 slovami a tri snímky overovací súbor s 1 121 slovami. V prípade tohto variantu sa úroveň chybovosti znížila na hodnotu 3,26 %. Najúspešnejší spomedzi nich bol tretí pokus s výslednou hodnotou úrovne chybovosti 3,08 %. Oproti druhému pokusu som snímky doplnila o základný model pre nemčinu „StAZH_RRB_German_Kurrent_XIX“. Teda pri navýšení počtu snímok v modeli len o štyri s 841 slovami došlo k zdokonaleniu existujúceho modelu 44919 o desať stotín percenta.



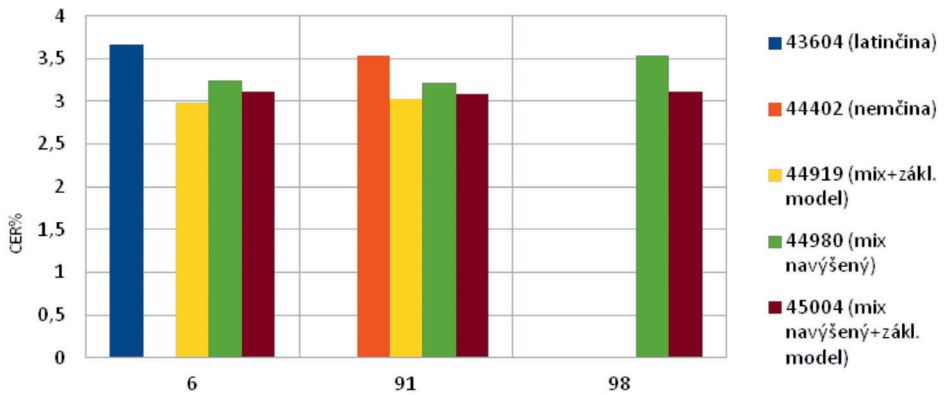
Obrázok 51 Graf modelu 45004. Zdroj: *Transkribus*.

Konečný model s identifikačným číslom 45004 bol otestovaný na výbere štyroch snímok druhej tematickej skupiny katalógu (Columna II. Continet Patrologiam, & Theologiam Dogmaticam, eamque a) Scholasticam & b) Catechet.). Úspešnosť modelu sa potvrdila, keď sa miera chybovosti nových snímok pohybovala v rozmedzí 2,05 % – 3,23 % (snímka č. 11: 3,23 %, snímka č. 12: 2,05 %, snímka č. 18: 3,22 %, snímka č. 19: 2,31 %). Vyššie percentuálne rozdiely v prepise jednotlivých snímok sú spôsobené špecifickými znakmi každej jednej z nich. Napríklad vyššia hodnota chybovosti v snímke č. 11 bola spôsobená zrejme zastúpením názvu tematickej skupiny, ktorý je oproti trénovanému textu odlišný. Naopak nižšiu hodnotu chybovosti v prípade snímky č. 12 je možné vysvetliť častejšie sa vyskytujúcou skratkou „detto“, ale hlavne použitím latinky v celom jej texte.

Zo zhodnotenia oboch fáz výskumu konfrontáciou samostatných modelov oproti spoločnému modelu vyplýva, že spoločný model dvoch typov písma dosiahol porovnateľné výsledky voči latinskému humanistickému písmu a zlepšenie voči nemeckému kurentu. Konkrétne hodnoty sa pohybujú v závislosti od jednotlivých snímok.

Tabuľka 10 Prehľad % chybovosti konkrétnych snímok v jednotlivých modeloch.

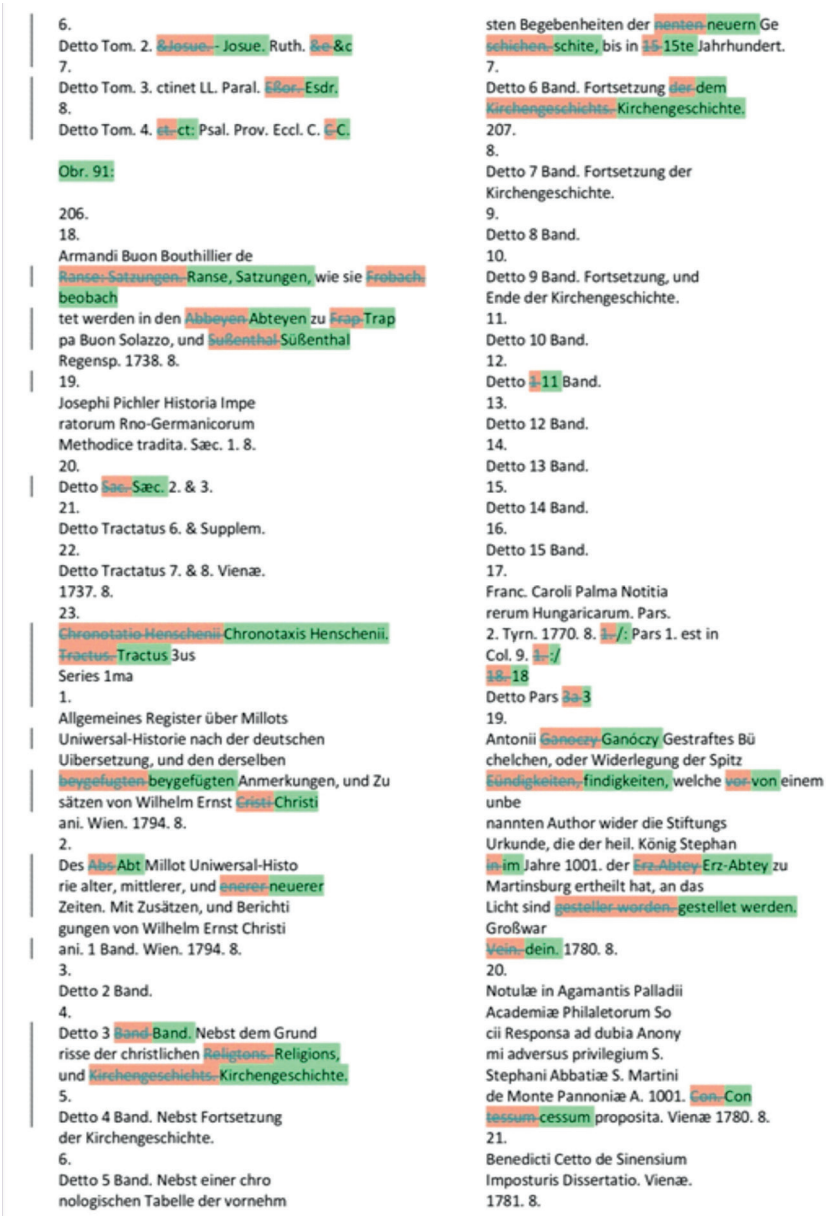
Model č.	ID	snímka č.			CER % modelu
		6	91	98	Overovací súbor
3	43604	3,66	x	x	3,05
6	44402	x	3,54	x	3,39
9	44919	2,98	3,03	x	3,18
11	44980	3,24	3,22	3,53	3,26
12	45004	3,11	3,08	3,11	3,08



Obrázok 52 Grafické vyjadrenie % chybovosti.

Aj napriek zlepšeným hodnotám jednotlivých modelov pravidelne sa v nich vyskytujú niektoré typy chýb. Pri vyhodnocovaní znakov, ktoré boli nesprávne prepísané, mali najväčšie zastúpenie chyby v oblasti interpunkcie a diakritiky. Patria do oblasti slabších nedostatkov, ktoré nemajú takmer žiadny vplyv na zrozumiteľnosť transkribovaného textu. Rovnako väčší podiel mali aj nedostatky spôsobené nepravidelným riadkovaním, čo malo za následok presah znakov. Konkrétne horné a dolné časti jednotlivých znakov

sa prekrývajú s ďalšími riadkami. Tieto chyby sa radia medzi závažnejšie, ale ich výskyt by sa pravdepodobne neznížil ani nasledujúcimi tréningmi. Ďalší okruh chýb vyplynul čiastočne už z tréningu modelu v procese vlastného prepisu textu. Môžeme sem zaradiť zámeny písmen, jednak v ich forme pre malé a veľké písmená, jednak pre podobné tvary písiem (príklad pre kurent e/r).



6.
Detto Tom. 2. &Josue. - Josue. Ruth. &c &c
7.
Detto Tom. 3. ctinet LL. Paral. Eßor. Esdr.
8.
Detto Tom. 4. et. ct: Psal. Prov. Eccl. C. C.

Obr. 91:

206.
18.
Armandi Buon Bouthillier de
Ranse: Satzungen. Ranse, Satzungen, wie sie Frobach.
beobach
tet werden in den Abbeyen Abteyen zu Frap Trap
pa Buon Solazzo, und Sußenthal Süßenthal
Regensp. 1738. 8.
19.
Josephi Pichler Historia Impe
ratorum Rno-Germanicorum
Methodice tradita. Sæc. 1. 8.
20.
Detto Sæc. Sæc. 2. & 3.
21.
Detto Tractatus 6. & Supplem.
22.
Detto Tractatus 7. & 8. Vienæ.
1737. 8.
23.
Chronotatio Henschenii Chronotaxis Henschenii.
Tractus. Tractus 3us
Series 1ma
1.
Allgemeines Register über Millots
Uniwersal-Historie nach der deutschen
Uibersetzung, und den derselben
beygefugten beygefügten Anmerkungen, und Zu
sätzen von Wilhelm Ernst Cristi Christi
ani. Wien. 1794. 8.
2.
Des Abs Abt Millot Uniwersal-Histo
rie alter, mittlerer, und enerer neuerer
Zeiten. Mit Zusätzen, und Berichti
gungen von Wilhelm Ernst Christi
ani. 1 Band. Wien. 1794. 8.
3.
Detto 2 Band.
4.
Detto 3 Band Band. Nebst dem Grund
risse der christlichen Religtons. Religions,
und Kirchengeschichts. Kirchengeschichte.
5.
Detto 4 Band. Nebst Fortsetzung
der Kirchengeschichte.
6.
Detto 5 Band. Nebst einer chro
nologischen Tabelle der vornehm
sten Begebenheiten der henten neuern Ge
schichen. schite, bis in 15. 15te Jahrhundert.
7.
Detto 6 Band. Fortsetzung der dem
Kirchengeschichts. Kirchengeschichte.
207.
8.
Detto 7 Band. Fortsetzung der
Kirchengeschichte.
9.
Detto 8 Band.
10.
Detto 9 Band. Fortsetzung, und
Ende der Kirchengeschichte.
11.
Detto 10 Band.
12.
Detto 1. 11 Band.
13.
Detto 12 Band.
14.
Detto 13 Band.
15.
Detto 14 Band.
16.
Detto 15 Band.
17.
Franc. Caroli Palma Notitia
rerum Hungaricarum. Pars.
2. Tyrn. 1770. 8. 1. /: Pars 1. est in
Col. 9. 1. :/
18. 18
Detto Pars 2a. 3
19.
Antonii Ganoczy Ganóczy Gestraftes Bü
chelchen, oder Widerlegung der Spitz
Eündigkeiten, findigkeiten, welche vor von einem
unbe
nannten Author wider die Stiftungs
Urkunde, die der heil. König Stephan
in im Jahre 1001. der Erz-Abtey Erz-Abtey zu
Martinsburg ertheilt hat, an das
Licht sind gestellet worden. gestellet werden.
Großwar
Vein. dein. 1780. 8.
20.
Notulæ in Agamantis Palladii
Academiæ Philaletorum So
cii Responsa ad dubia Anony
mi adversus privilegium S.
Stephani Abbatiæ S. Martini
de Monte Pannoniæ A. 1001. Con. Con
tessum cessum proposita. Vienæ 1780. 8.
21.
Benedicti Cetto de Sinensium
Imposturis Dissertatio. Vienæ.
1781. 8.

Obrázok 53 Ukážka porovnania automatickej transkripcie snímky č. 91 (model 45004) s jeho korigovanou verziou. Zdroj: *Transkribus*.

ZÁVER

Napriek tomu, že rukopisný katalóg knižnice nespĺňa všetky zavedené štandardy pre prácu s platformou *Transkribus*, neznižuje to jeho možnosti na vytvorenie funkčného modelu automatickej transkripcie. Menší počet tréningových údajov účinne kompenzuje správne zvolený základný model. Nástroje platformy sú schopné efektívne rozlíšiť dva výrazne odlišné typy písma v jednom modeli, a to za využitia len 20 % z celkového rozsahu textu prameňa.

Pozitívne hodnotenie si zasluhuje okrem iného aj vo všeobecnosti dobrý automatický prepis číselných znakov, keďže sa v dokumente vyskytujú opakovane v podobe čísel strán, záznamov, ako i vo forme rokov vydania a formátov kníh. Časovú náročnosť práce naopak zvyšuje hlavne manuálna tvorba tabuliek a potreba dôslednej kontroly segmentácie riadkov.

Dobrý výsledok ovplyvnil ale aj samotný prameň, pretože ho tvorí úhľadné jednoliate písmo a jeho text zahrnutý do tréningového procesu vznikol jednou pisárskou rukou v krátkom časovom období. Vplyv mal aj dobrý fyzický stav rukopisu, ktorý až na menšie atramentové machule a občasné presvitajúce písmo nenesie výraznejšie znaky poškodenia papiera.

Získaný model s použitím nástroja HTR+ je schopný správne prečítať približne 97 % údajov z požadovaného textu rukopisného katalógu, čo predstavuje veľmi dobrú funkčnosť automatickej transkripcie. Na druhej strane treba pripomenúť aj to, že malý model má svoje obmedzenia, je veľmi dobre prispôsobený na konkrétny rukopis, ale nie je možné ho použiť neuvážene na iné rukopisy.

Praktické využitie modelu je z toho dôvodu prioritne zamerané na automatickú transkripciu celého rukopisu prvého katalógu. Pre samotný prameň je však významný aj iný nástroj platformy *Transkribus* v podobe metadátového vyhľadávania. Preto ďalším krokom v jeho výskume bude tvorba metadát, ktorými budú označené jednotlivé textové rámce so základnými prvkami (napr. meno autora, miesto vydania, rok vydania) potrebnými na následnú podrobnú analýzu fondu knižnice.

Druhá línia výskumu bude smerovať k zdokonaľovaniu samotného modelu, či už zaradením podobných rukopisov z doplnkového stĺpca prvého katalógu do tréningového procesu, ale aj pokusmi s rukopismi zvyšných častí prameňa. Funkčnosť modelu overia aj ďalšie zachované rukopisné pramene knižnice. V Literárnom archíve Slovenskej národnej knižnice evidujú napríklad zlomok knižničného katalógu banskobystrického diecézneho seminára z obdobia okolo roku 1803, ktorý by mohol niesť totožný rukopis ako skúmaný prvý katalóg knižnice.

Kňazský seminár sv. Karola Boromejského sa počas svojej existencie stal významnou pedagogickou inštitúciou na pôde mesta Banskej Bystrice. Vysokú úroveň v oblasti prípravy seminaristov na budúce kňazské povolanie možno pripísať, samozrejme, dobrej erudovanosti viacerých profesorov, organizačnej aktivite vedenia, ale taktiež starostlivo vybudovanej a bohatej knižnici. K jej prehľadnosti prispievali aj knižničné katalógy. Sú to neoceniteľné pomôcky, ktoré dnes ponúkajú rozsiahle množstvo informácií, a aj preto sa stali dobrou príležitosťou na využitie moderných technológií, ktoré nenahrádzajú prácu historikov, ale im pomáhajú.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

200 rokov Kňazského seminára v Banskej Bystrici: 1807 – 2007. Banská Bystrica ; Badín: Kňazský seminár sv. Františka Xaverského. 155 s.

AGNET, Ján: Historické knižné fondy na Slovensku: (genéza, vývoj a súčasný stav). In: *Teória a výskum knihovníctva a bibliografie. Výskumy č. 40: historické knižničné fondy na Slovensku*. Martin : Matica slovenská, 1988, s. 17 – 30.

DECKER, Viliam: *Dejiny ručnej výroby papiera na Slovensku*. Martin : Matica slovenská, 1982. 224 s.

Diecézny archív v Banskej Bystrici, fond Kanonické vizitácie, sign. CV 24.

GYÖRGY, Aladár (ed.): *Hivatalos statisztikai közlemények: magyarsrszág köz- és magánkönytárai 1885-ben*. Budapest : Athenaeum, 1886.

JANKOVIČ, Vendelín: *Rukopisy Miestneho pracoviska Matice slovenskej v Bratislave*. Matin : Matica slovenská, 1958.

KONIAROVÁ, Anna: *Dejiny banskobystrickej diecézy v 18. a 19. storočí*. Badín : Kňazský seminár sv. Františka Xaverského, 2002. 185 s.

KUZMÍK, Jozef: Historické knižnice na Slovensku a ich postavenie v našej kultúrnej minulosti a prítomnosti. In: *Letopis pamätníka slovenskej literatúry*, 1968, s. 165 – 192.

Magyar minerva a magyarországi múzeumok és könyvtárak címkönyve. IV. évfolyam: 1904 – 1911. Budapest : Athenaeum, 1912.

Magyar minerva a magyarországi múzeumok és könyvtárak címkönyve V. kötet: 1912 – 1913. Budapest : Athenaeum, 1915.

MÉSZÁROSOVÁ, Klára: Katalóg jezuitskej knižnice v Banskej Bystrici z roku 1778. In: *Kniha 95 – 96. Zborník o problémoch a dejinách knižnej kultúry*. zost. M. Domová, Martin : Matica slovenská, 1997, s. 148 – 157.

MIŠÍK, Mikuláš: Banskobystrickí bohoslovci v slovenskom národnom živote. In: *Zborník literárno-vedeckého odboru Spolku sv. Vojtecha*, tom. 2, vol. 2, 1935, s. 9 – 153.

Ratio educationis 1777 a 1806: prvá jednotná sústava výchovy a vzdelávania v dejinách našej kultúry. Bratislava : Slovenské pedagogické nakladateľstvo, 1988. 454 s.

SABOV, Peter: Historické knižničné fondy v správe SNK Matice slovenskej. In: *Teória a výskum knihovníctva a bibliografie. Výskumy č. 40: historické knižničné fondy na Slovensku*. Martin : Matica slovenská, 1988, s. 33 – 47.

Štátna vedecká knižnica v Banskej Bystrici, fond starých a vzácnych dokumentov, sign. MS Ba H 125.

TATÁRIKOVÁ, Monika: Katalóg knižnice banskobystrického biskupského seminára sv. Karola Boromejského zo začiatku 19. storočia. In: *Kniha 2019. Zborník o problémoch*

a dejinách knižnej kultúry. zost. D. Škulová, Martin : Slovenská národná knižnica, 2019, s. 200 – 209.

Zoznam rukopisných katalógov historických knižníc na Slovensku. Martin : Knihovedné stredisko Matice slovenskej, 1959.

KAPITOLA 8

MODELY AUTOMATICKEJ TRANSKRIPCIE ŠTVORJAZYČNÉHO DIELA J. A. KOMENSKÉHO ORBIS PICTUS (1798)

Lucia Nižníková – Michaela Mikušková

Univerzita Mateja Bela v Banskej Bystrici; Univerzitná knižnica

E-mail: lucia.niznikova@umb.sk; michaela.mikuskova@umb.sk

ABSTRAKT

Možnosti automatickej transkripcie tlačeného dokumentu pomocou softvéru T*ranskribus* sme overovali na štvorjazyčnom vydaní diela J. A. Komenského *Orbis Pictus* z roku 1798. Špecifikom dokumentu je aj to, že pri tlači boli použité štyri typy fontu – antikva, kurzíva, fraktúra a švabach. Dobový stav jazykov reflektuje používanie grafém, ktoré sa dnes už nepoužívajú. Z tohto dôvodu sme ako metódu prepisu zvolili transliteráciu, ktorá umožňuje spätnú rekonštrukciu pôvodnej podoby slov a zachovanie vysokej autenticity prepísaného originálneho textu. Na trénovanie modelov sme použili tri metódy. Každou z nich sme dosiahli relatívne dobré výsledky, ktoré sa podľa dostupných štúdií dajú očakávať pri historických tlačiach. Celkovo deväť modelov sme trénovali pomocou technológií HTR+ a PyLaia. Najúspešnejší model v ďalšej fáze využijeme na automatickú transkripciu celého dokumentu, ktorý následne doplníme o textové metadáta, metadáta na úrovni dokumentu a redakčné vyhlásenie. Finálny dokument bude sprístupnený odbornej a laickej verejnosti na platformách softvéru *Transkribus*.

Kľúčové slová: automatická transkripcia, transliterácia, trénovanie modelov, strojové učenie, platforma *Transkribus*, *Orbis Pictus*

ABSTRACT

Automatic transcription models of the four-language work Orbis pictus (1798) by J. A. Comenius

We verified the possibilities of automatic transcription of a printed document using the *Transkribus* software on the four-language edition of the work *Orbis Pictus* by J. A. Comenius from 1798. A specific feature of the document is that four font types were used for printing – Antikva, Italic, Fracture and Schwabach. The period state of languages reflects the use of graphemes that are no longer used today. For this reason, we chose transliteration as the method of transcription, which enables the backward reconstruction of the original form of the words and the preservation of the high authenticity of the transcribed original text. We used three methods for model training. With each of them, we achieved relatively good results, which, according to

the accessible studies, can be expected with historical prints. We trained a total of nine models using HTR+ and PyLaia technologies. In the next phase, we will use the most successful model for the automatic transcription of the entire document, which will be subsequently supplemented with text metadata, document-level metadata and an editorial statement. The final document will be made accessible to the professional and lay public on the *Transkribus* software platforms.

Keywords: automatic transcription, transliteration, model training, machine learning, *Transkribus* platform, *Orbis Pictus*

ÚVOD – HISTORICKÉ POZADIE

Osvietenské vládnutie Márie Terézie sa vo výraznej miere prejavilo vo všetkých oblastiach spoločenského života, školstvo nevynímajúc. Reformné opatrenia týkajúce sa správy a vnútornej náplne škôl prijaté v rakúskej monarchii v roku 1774 sa v Uhorsku odzrkadlili o tri roky neskôr prijatím organizačnej správy pre školskú sústavu Ratio Educationis (1777). Tieto reformy presadzujúce modernizáciu obsahu a metód školského vyučovania možno považovať za najvýznamnejšie dokumenty v oblasti vývoja ľudového školstva. Vychádzajú z potreby zvýšiť rozsah vzdelanosti prostého ľudu zavedením povinnej školskej dochádzky, priblížiť ho praktickým potrebám ľudu a oslabiť vplyv cirkvi na názory ľudí v oblasti vedy. Jednotné organizovanie a centralizovanie školstva bolo spojené aj s vydávaním nových učební. V 80. rokoch 18. storočia sa skončilo obdobie tvrdej cenzúry, ktorá prešla do rúk štátu. Kníhtlač sa stala nástrojom osvety, kultúry a vzdelanosti. Prestali sa prideľovať privilégiá na tlač učební, ktoré boli výhodným obchodným artiklom a žiadanou zložkou kníhtlačiarskej produkcie. Jazykový štýl učební vydávaných po reformách nebol ustálený. Bernolákova kodifikácia spisovnej slovenčiny vychádzajúca z potreby jednotnej normy jazyka sa prejavila aj v elementárnych učebniciach ľudového školstva. Ako uvádza Kowalská[1], učebnice po roku 1790 sa pridržiavali spisovnej normy, najmä tlače z produkcie trnavskej kníhtlačiarne, ale nestretli sa s porozumením evanjelickej časti používateľov. Okrem požiadavky na výučbu nemeckého jazyka sa v 90. rokoch 18. storočia objavuje aj potreba rozširovania maďarského jazyka. Vo vydávaní učební sa výrazne prejavilo aj používanie rôznych dialektov, najmä panónsko-slovenského dialektu, ktorý bol v rozpore s kodifikovanou bernolákovčinou používajúcou bohemizmy.

V období zavádzania reforiem mal na Slovensku výrazný vplyv pedagogický odkaz a pôsobenie Jana Amosa Komenského. Na šírení jeho diel v slovenskom prostredí mali výrazný podiel českí exkulanti a hlásili sa k nemu najmä evanjelické školy a učitelia. V roku 1633 na Slovensku prvýkrát vychádza kniha *Praxis pietatis*, šírenie Komenského myšlienok podporil aj jeho pobyt v Blatnom Potoku. V nasledujúcom období jeho učebnice vychádzali častejšie a našli uplatnenie nielen na školách, ale aj v bežnom

1 KOWALSKÁ, Eva: Učebnice pre štátne ľudové školy na Slovensku koncom 18. storočia. In: *Kniha ´90. Zborník o problémoch a dejinách knižnej kultúry na Slovensku.* zost. M. Domová – R. Brož, Martin : Matica slovenská, 1990, s. 72.

živote.[2] Pedagogický odkaz „učiteľa národov“ a mnohé pokrokové myšlienky v oblasti didaktiky, ktoré uplatnil vo viacerých dielach, pretrvávajú dodnes. Za základné princípy vedúce k poznaniu považoval zapojenie zmyslov, názornosť, primeranosť veku, prechod od jednoduchého k všeobecnému a permanentné vzdelávanie po celý život. Za vrcholné Komenského dielo sa považuje *Orbis Sensualium Pictus* (Svet v obrazoch), v ktorom uplatnil všetky svoje pedagogické názory a didaktické princípy vo vyučovaní jazykov. Zároveň je naplnením Komenského predstavy o obrazovej učebnici, ktorá sa dnes považuje aj za najstaršiu detskú encyklopédiu.

Orbis *Sensualium Pictus* sa v 70. rokoch 18. storočia na základe školského nariadenia začal používať ako bežná učebnica aj na katolíckych školách v rakúskych a uhorských krajinách.[3] Vychádzal v rôznych jazykových mutáciách a ich kombináciách. Učebnica pôvodne slúžila na vyučovanie latinčiny pre mladších žiakov. Postupom času bola viackrát prepracovaná, doplňovaná podľa nových vedeckých objavov a teórií a prispôsobovaná školským nariadeniam a učebným osnovám. Do roku 1842 vyšlo na Slovensku trinásť vydaní diela Orbis Pictus[4], najčastejšie v štyroch jazykoch (latinskom, maďarskom, nemeckom a českom). Prvé vydania pochádzajú z kníhtlačiarne Samuela Brewera v Levoči v rokoch 1685, 1728 a 1778. Od roku 1798 sa tlač diela Orbis Pictus presunula do Bratislavy, ktorá sa stala významným vydavateľským miestom Komenského diel. Prvým bratislavským kníhtlačiarom, ktorý sa v rokoch 1798 a 1820 zaslúžil o vydanie tejto učebnice, bol Šimon Peter Weber[5]. V Bratislave pôsobil od roku 1783, tlačiareň sa nachádzala na Michalskej ulici. Tlače z jej produkcie boli na veľmi dobrej typografickej úrovni. Vďaka bohatej zásobe písma a dobrého strojového zariadenia sa jej podarilo uspieť aj vo veľkej konkurencii bratislavských kníhtlačiarov (Ján Michal Landerer, František Augustín Packo, Anton Loewe). Z jeho dielne vychádzali tlačoviny v latinskom, nemeckom, maďarskom, českom a slovenskom jazyku. Z hľadiska grafickej úpravy a ilustrácií uplatňoval najmä výzdobu drevorezom a drobnými vinetovými motívmi. Určitý čas spolupracoval aj s Jánom Matejom Korabinským a v roku 1786 vydal jeho *Zemepisno-historický lexikon (Geographisch-Historisches und Produkten Lexikon von Ungarn).*

Základom diela je noetický senzualizmus a z neho vyplývajúce didaktické zásady, pedagogický realizmus a myšlienka pansofie, ktorej otcom je práve J. A. Komenský. Orbis Pictus vychádza z názorného zobrazenia sveta a aktuálnych vedeckých poznatkov v danom čase. Pôvodne učebnica latinčiny, neskôr aj iných jazykov, človeka vníma ako homo ludens, teda bytosť, ktorá svoje schopnosti a možnosti rozvíja pomocou hry.

2 MÁTEJ, Jozef: Pedagogické dedičstvo J. A. Komenského na Slovensku. In: *Acta Facultatis Philosophicae Universitatis Šafarikanae Prešovensis, Pedagogika II/1971*. red. F. Karšai, Košice : Východoslovenské vydavateľstvo v Košiciach, 1971, s. 27 – 45.

3 BOKROSOVÁ, Katarína: Zbierka Komenián zo 17. až 19. storočia v študovni historických tlačí Slovenskej pedagogickej knižnici v Bratislave. In: *Bibliotheca Antiqua* [online]. Olomouc: Vědecká knihovna v Olomouci, 2014 [cit. 2022-11-14]. Dostupné na: https://vmdelta.vkol.cz/data/soubory/hf/bibliotheca-antiqua-14/03bokrosova%20.pdf

4 BAKOŠ, Ľudovít. Didaktické aspekty Komenského „Orbis Pictus”. In: *Acta Facultatis Philosophicae Universitatis Šafarikanae Prešovensis, Pedagogika II/1971*. red. F. Karšai, Košice: Východoslovenské vydavateľstvo v Košiciach, 1971, s. 143.

5 CESNAKOVÁ-MICHALCOVÁ, Milena. Šimon Peter Weber: portrét bratislavského kníhtlačiara a dramatika. In: *Slovenské divadlo: revue dramatických umení*, roč. 24, č. 4, 1976, s. 571 – 587.

Komenský bol v Európe prvý, kto presadzoval zásadu škola hrou, keď v roku 1654 zostavil súbor školských divadelných hier s názvom Schola ludus[6] v snahe oživiť výučbu latinčiny hrou – vystupovaním žiakov na javisku. Orbis Pictus možno vnímať ako návod pre učiteľov, že majú vyučovať názorne a v prírode alebo v prostredí, ku ktorému sa výučba tematicky viaže.

Vydanie z roku 1798 spolu v 82 tematických kapitolách pokrýva základy spoločenských a prírodných vied, náboženstvo, remeslá, domáce hospodárstvo, poľnohospodárstvo i reálie, ktoré odzrkadľujú každodenný život (trestanie zločincov, pohreb). Každá kapitola obsahuje pojmy viažuce sa k danej téme tak, aby neboli vnímané izolovane, ale vo vzájomnom pôsobení a vzťahoch. Počet pojmov v rámci kapitol sa líši, v tomto vydaní je to 3 až 30 pojmov v jednej kapitole. Texty pútavým spôsobom dopĺňajú názorné ilustrácie. Tvoria ich súhrnné obrázky, v ktorých sú jednotlivé predmety označené číslicami. Tie ich prepájajú s pojmami vysvetlenými pod ilustráciou. Komenský vychádza zo svojich pedagogických skúseností, kladie dôraz na názorné pochopenie vecí a ich obrazu, uprednostňuje význam pojmov pred gramatikou a princíp paralelizmu jazykového vyučovania. Domnieva sa, že žiaci by nemali naučenú látku len mechanicky odriekať, ale v prvom rade rozumieť tomu, čo sa učia. Kladie dôraz na zásady pansofie, systematického a zrozumiteľného zhrnutia základných poznatkov a pojmov. Učebnica je spracovaná tak, aby bola zrozumiteľná aj malému dieťaťu a zároveň poskytovala ľahkú a príjemnú cestu vecného a jazykového učenia sa prostredníctvom názornosti a obsahového prepojenia.

FORMÁLNY POPIS DOKUMENTU

Historická tlač vydaná v kníhtlačiarni Šimona Petra Webera v Bratislave v roku 1798 je súčasťou historického knižničného fondu Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici. Stoosemdesiatštyristranová publikácia s rozmermi 18,9 x 11,5 cm je vytlačená na ručne odlievanom papieri s čiastočnými priesvitkami v tvare slnka a iniciálou L. Tento fragment nezodpovedá žiadnej zdokumentovanej priesvitke zo slovenských ručných papierní v publikácii Viliama Deckera[7], preto sa dá predpokladať, že Weber dovážal papier zo zahraničia. Tlač je viazaná v lepenkovej väzbe s pôvodným dekoratívnym papierom. Textovej časti predchádza frontispis, ktorý zobrazuje hospodárske a lesné zvieratá v ich prirodzenom prostredí. V hornej časti frontispisu je umiestnený erb Uhorska. Titulný list je podobne ako zvyšný text publikácie štvorjazyčný bez výrazných tlačiarenských ozdôb. Číslovanie strán je priznané na stranách jednotlivých kapitol, na frontispise, titulnom liste a stranách jazykových indexov sa nenachádza. Paginácia je zdobená malými typografickými ozdobami s motívom kvetu. Poradie strán určujú aj kustódy. Poslednú kapitolu uzatvára okrúhla vineta s podobizňou leva a priemerom 3,6 cm. V textovej časti sú početne zastúpené

6 PŠENÁK, Jozef: Jan Amos Komenský, tvorca školských divadelných hier. In: *Paedagogica. Zborník Filozofickej fakulty Univerzity Komenského*, roč. 17. red. Š. Švec, Bratislava : Univerzita Komenského v Bratislave, 2005, s. 134.

7 DECKER, Viliam. *Dejiny ručnej výroby papiera na Slovensku*. Martin : Matica slovenská, 1982. 223 s.

obdĺžnikové drevoryty vo veľkosti 8,5 x 5,4 cm s výnimkou druhej kapitoly, v ktorej sa nachádzajú dva okrúhle drevoryty s priemerom 5,1 cm. Cieľom sprievodných ilustrácií je čo najvyššia autentickosť a názornosť. Okrem obrázkového motívu viažuceho sa na obsah kapitoly sa v drevorytoch nachádza číslovanie pojmov vysvetlených v texte. Autorstvo drevorytov je otázne, pretože nie sú signované. Domnievame sa, že ich autormi môžu byť rytci cudzieho pôvodu. Ako uvádza Šturdíková-Hudáková[8], Š. P. Weber spolupracoval najmä s Marcusom Weinmannom, ktorý v Bratislave pôsobil v 80. a 90. rokoch 18. storočia.

Prvé vydanie diela *Orbis Sensualium Pictus*[9] z roku 1658 v Norimbergu bolo dvojjazyčné, latinsko-nemecké, v dvojstĺpcovej grafickej úprave s obrázkami z drevorytov Paula Creutzbergera.[10] Prvé štvorjazyčné latinsko-maďarsko-nemecko-české vydanie vyšlo v roku 1685, jeho český názov znel *Svět Wyditedlny*.[11] Prvé štvorjazyčné vydanie z dielne prešporského tlačiara Š. P. Webera z roku 1798 má český názov *Svět Namalowaný*.[12]

Text je formálne usporiadaný do štyroch blokov. Latinský, maďarský a nemecký v stĺpcoch so zachovaním medzier tak, aby boli jednotlivé jazykové verzie súvisiace s rovnakými pojmami na jednej úrovni. Česká mutácia textu má klasickú knižnú formu cez celú stranu. Súčasťou dokumentu sú abecedné registre základných pojmov v latinčine, maďarčine, nemčine a češtine.

Každý jazyk je vytlačený iným fontom – latinčina antikvou, maďarčina kurzívou, nemčina fraktúrou a čeština švabachom, o to je dokument zaujímavejší hľadiska skúmania možností automatickej transkripcie pomocou softvéru *Transkribus*.

Na titulnom liste sa nachádzajú dva provenienčné záznamy vo forme pečiatky, jeden rukopisný posesorský zápis, inventarizačná pečiatka a rukou napísaná signatúra. Staršia pečiatka patrí Evanjelickému lýceu v Banskej Štiavnici (SELMECI EVANG. LYCEUM), ktoré na území mesta v rôznych podobách ako nižšia evanjelická škola, neskôr nižšie a úplné gymnázium a napokon dištriktuálne gymnázium s podporou banského dištriktu pôsobilo od roku 1587 do roku 1919, keď z dôvodu zastavenia štátnej podpory ukončilo svoju činnosť. Pečiatka sa nachádza aj na stranách 58 a 95. K pečiatke patrí rukou napísaná signatúra 38.

8 ŠTURDÍKOVÁ-HUDÁKOVÁ, Marta: Bratislavskí tlačiari 18. storočia a ich vyzdobené tlače. In: *Kniha ´82. Zborník pre problémy a dejiny knižnej kultúry na Slovensku*. zost. J. Valach, Martin : Matica slovenská, 1983, s. 109.

9 COMMENII, Joh. Amos: *Orbis Sensualium Pictus*. Noribergae : Typis & Sumptibus Michaelis Endteri, 1658. Zdigitalizované dostupné na: http://diglib.hab.de/drucke/47-7-eth-as/start.htm

10 COMENIUS, Iohannes Amos. Orbis Sensualium Pictus. nem. preklad Siegmund von Birken, ilustr. Paulus Creutzberger. In: *Bibliotheca Augustana* [online]. [cit. 2022-11-20]. Dostupné na: http://www.hs-augsburg.de/~harsch/Chronologia/Lspost17/Comenius/com_o000.html

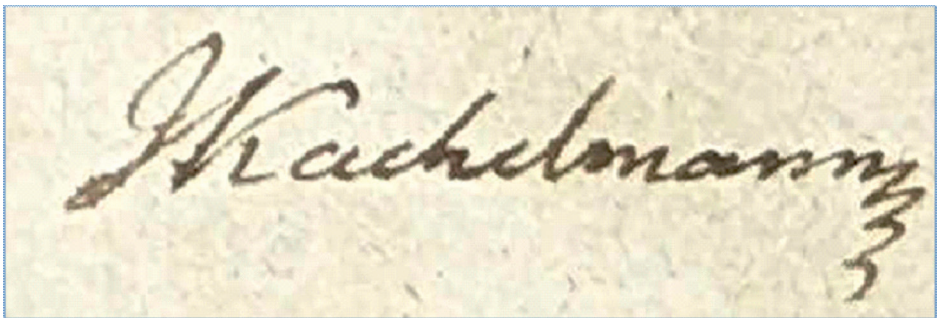
11 COMENII, Joh. Amos: *Orbis Sensualium Pictus Quadrilinguis*. Leutschoviae : Typis Samuelis Brewer, 1685. Zdigitalizované dostupné na: https://www.slovakiana.sk

12 COMENII, Ioann. Amos: *Orbis Pictus, in hungaricum, germanicum et slavicum translatus, et hic ibive emendatus*. Posonii : Sumtibus & Typis Simonis Petri Weber, 1798. Zdigitalizované dostupné na: https://www.slovakiana.sk/

Obrázok 54 Pečiatka Evanjelického lýcea v Banskej Štiavnici a ručne napísaná signatúra Zdroj: Orbis Pictus (1798).

Knižnica evanjelického lýcea krátko pred zánikom školy obsahovala takmer 18 000 zväzkov kníh. Po zániku zostal fond knižnice vlastníctvom evanjelickej a. v. cirkvi v Banskej Štiavnici. Začas sa uchovával v budove lýcea, neskôr bol viackrát presťahovaný. Od roku 1996 je deponovaný a uložený v budove Slovenského banského múzea v Banskej Štiavnici. Podľa dostupných informácií bol knižničný fond budovaný kúpou a tiež knižnými darmi. Svoje knižné zbierky venovali knižnici viacerí profesori lýcea, ako napr. Ľudovít Matej Šuhajda, Ján Brezník, Ján Belluš, Jozef Paulovič a iní. Knihami a peňažnými čiastkami prispievali na rozvoj knižnice viacerí priaznivci školy, obyvatelia Banskej Štiavnice ako lekár Štefan Bolleman či právnik a advokát Ján Kachelmann

V historickom katalógu lyceálnej knižnice boli zistené tieto údaje: kniha bola zaradená do predmetovej skupiny Ph – philosophia. Mala priradenú signatúru Ph 38, prírastkové číslo 6898. V prírastkovom zozname nie je uvedený presný dátum ani rok prírastku, uvedený spôsob nadobudnutia je dar (1 zväzok), cena knihy 1 koruna 80 filérov. Meno darcu nie je zapísané v katalógu, v poznámke ani na pôvodnom historickom katalógovom lístku – podľa rukopisného záznamu v knihe by však malo ísť o Jána Kachelmanna.[13]



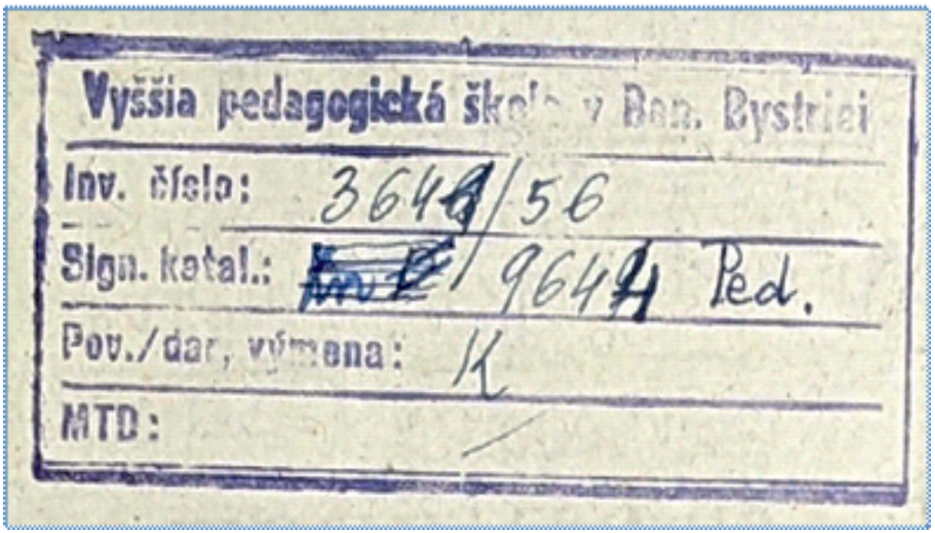
Obrázok 55 Rukopisný posesorský zápis. Zdroj: Orbis Pictus (1798).

Provenienčný záznam Dekanstva Vyššej pedagogickej školy v Banskej Bystrici na titulnom liste sa nachádza aj na s. 184. K nemu patrí inventarizačná pečiatka so signatúrou 9644.

13 PAVÚK, Marián: Kachelmannovci. In: *Haló Vyhne* [online], č. 3, 2003, s. [1 – 2] [cit. 2022-10-24]. Dostupné na: https://www.vyhne.sk/kachelmannovci/mid/66591/.html#m_66591

Obrázok 56 Pečiatka Dekanstva Vyššej pedagogickej školy v Banskej Bystrici a signatúra. Zdroj: Orbis Pictus (1798).

Na rube titulného listu sa nachádza podrobná inventarizačná pečiatka Vyššej pedagogickej školy v Banskej Bystrici s inventarizačným číslom 3641/56. Podľa zápisu v prírastkovom zozname a údajov na tejto pečiatke dokument získala Vyššia pedagogická škola v roku 1956 formou kúpy v antikvariáte na území mesta Banská Bystrica. V prírastkovom zozname je vedená pod číslom 3641. Mala priradenú signatúru 9644, bola zaradená do predmetovej skupiny Ped – pedagogika. Knižnica Vyššej pedagogickej školy v Banskej Bystrici sa v roku 1992 stala základom fondu Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici.[14]



Obrázok 57 Inventarizačná pečiatka Vyššej pedagogickej školy v Banskej Bystrici. Zdroj: Orbis Pictus (1798).

14 HOMOLOVÁ, Ľudmila: *Sprievodca historickým knižničným fondom Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici*. Banská Bystrica : Univerzita Mateja Bela v Banskej Bystrici, 2012, 48 s.

VYHOTOVENIE DIGITALIZÁTU

Základ platformy tvorí program na automatické rozpoznávanie textu a súbory nástrojov na digitalizáciu, automatickú transkripciu, editovanie, plnotextové prehľadávanie a sprístupňovanie historických dokumentov. Primárne bol vyvinutý na prácu s rukopisnými textami a tvorbu HTR modelov (Handwritten Text Recognition), rovnako dobre však funguje aj pri spracovávaní tlačených dokumentov.[15] OCR technológia alebo optické rozpoznávanie znakov v sebe zahŕňa prvky umelej inteligencie a strojového čítania. Je schopná vytvoriť digitálnu, počítačom čitateľnú verziu vytlačeného alebo písaného dokumentu do formy použiteľnej a editovateľnej v textovom editore. Presnosť výstupu závisí najmä od zložitosti softvéru. Osobitnú kapitolu tvoria softvéry, ktoré sú schopné rozpoznávať rukou napísaný text. Tieto sú zvyčajne vybavené schopnosťou naučiť sa formu a štýl rukopisu a zohľadňovať pri tom meniaci sa sklon či veľkosť písma. Ide o technológiu strojového učenia (angl. machine learning), ktorá zrýchľuje a optimalizuje toto učenie pomocou špecializovaného procesora, tzv. neural engine. Jedným z príkladov takéhoto softvéru je aj *Transkribus*.

Na prácu s *Transkribom* je nevyhnutné mať k dispozícii dokument v digitálnej podobe. Podporované sú súbory vo formátoch JPEG, PNG, TIFF a PDF. Dokument, s ktorým sme pracovali, bol dostupný len v tlačenej forme, preto bolo potrebné vyhotoviť digitalizát – fotografické snímky pomocou mobilného telefónu a pomôcky na snímanie dokumentov vyvinutej na prácu s *Transkribom*, tzv. Scan Tentu.[16] S použitím mobilného telefónu Google Pixel 4 sme nasnímali celý dokument. Vyhotovili sme celkom 82 snímok zachytávajúcich dvojstranu originálneho dokumentu pri rozlíšení 72 DPI (veľkosť snímky 2 – 2,5 MB).

PRÍPRAVA VZORKY GROUND TRUTH

Po nahratí zdigitalizovaného dokumentu na platformu *Transkribus* nasledovala príprava vzorky Ground Truth. Tento pojem sa vo všeobecnosti používa v strojovom učení na označenie presných, objektívnych informácií poskytovaných empirickými, priamymi procesmi a nie informácií odvodených zo zdrojov prostredníctvom štatistického výpočtu neistoty. V prípade programu *Transkribus* sa informácie Ground Truth získavajú trénovaním s dostatočným množstvom údajov týkajúcich sa jednotlivého písma, ktoré sa potom môžu použiť na vytvorenie modelu, ktorý sa dá úspešne použiť na veľké objemy toho istého písma. Príprava vzorky pozostávala z dvoch krokov: segmentácie celého dokumentu na textové rámce, riadky a základné riadky (tzv. Baselines) a manuálneho prepisu strán vybraných do vzorky.

Segmentácia dokumentu predstavuje podrobné rozčlenenie štruktúry a orientácie textu v dokumente. Tento krok je nevyhnutný na správny prepis textu, jeho čítanie (poradie

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

16 KATUŠČÁK, Dušan: Digital Humanities a automatická transkripcia rukopisných textov. In: *ITLib*, roč. 24, č. 1, 2020, s. 10.

riadkov) a trénovanie modelu. Na tieto účely má *Transkribus* vyvinuté nástroje, ktoré umožňujú automatické rozpoznanie textových rámcov a jednotlivých riadkov v nich. Originál dokumentu historickej tlače Orbis Pictus obsahuje text v stĺpcoch, preto automatická segmentácia nebola možná. Z tohto dôvodu sme najprv manuálne vymedzili textové rámce a následne sme pristúpili k automatickej segmentácii základných riadkov.

Samostatne sme vymedzili rámce pre texty, ktoré neboli zoradené v stĺpcoch (čísla strán, nadpisy kapitol, kustódy a názvy kapitol v švabachu), a rámce textov usporiadaných v stĺpcoch (jazykové mutácie vytlačené antikvou, kurzívou a fraktúrou). Na segmentáciu textových rámcov v stĺpcoch bolo možné použiť dve metódy: využitie funkcie *Tabuľka* s možnosťou horizontálneho členenia stĺpcov a kopírovania nastaveného formátu na nasledujúce strany alebo manuálne vytváranie samostatných rámcov pre každý stĺpec. Využili sme obe metódy, kvalita segmentácie bola rovnaká. Následne bolo potrebné skontrolovať a v prípade potreby upraviť poradie čítania jednotlivých rámcov tak, aby za sebou logicky nasledovali.

Na manuálne preddefinované textové rámce sme spustili automatickú segmentáciu riadkov. Správna segmentácia riadkov má kľúčový vplyv na trénovanie modelu. Automatická segmentácia na malé výnimky zadefinovala a označila všetky riadky. V stĺpcovej časti textu sa vyskytli chyby pri identifikácii začiatku a konca riadkov. Systém nedokázal automaticky rozpoznať najmä prvé a posledné písmená riadka, rozdeľovníky a interpunkciu (bodky, čiarky, dvojbodky). Korekcie bolo nevyhnutné vykonať manuálne takmer na každom riadku. Znížila sa tým efektívnosť automatickej segmentácie, napriek týmto nedostatkom však bola užitočná. Domnievame sa, že chybovosť bola spôsobená tým, že ohraničenie textových rámcov sa nachádzalo v tesnej blízkosti riadkov. V textových rámcoch písma švabach, ktoré neboli súčasťou stĺpcovej časti textu, sa podobné chyby objavovali sporadicky. Túto tézu potvrdzuje vo svojej diplomovej práci aj Katreniak.[17]

Druhý problém, ktorý sa pri automatickej segmentácii vyskytol, boli spojené riadky medzi textovými rámcami. Opakoval sa v stĺpcovej časti textu a podobne ako chybné doťahovanie začiatkov a koncov riadkov súvisel s úzkymi okrajmi medzi čiarami stĺpcov a samotným textom. Spojené riadky presahujúce textové rámce bolo potrebné manuálne rozdeliť.

Pri automatickej segmentácii sa výnimočne niektoré riadky vôbec nenasegmentovali, preto sme ich dopĺňali manuálne. Naopak v niektorých prípadoch boli ako riadok označené nečistoty a text presvitajúci z opačnej strany listu.

Po oprave chýb z automatickej segmentácie bolo potrebné skontrolovať správne poradie riadkov. Chybné číslovanie poradia vzniklo vtedy, keď program po manuálnom rozdelení spojených riadkov v rôznych textových rámcoch nezaradil oddelené časti riadka do správneho textového rámca. Presun riadkov medzi textovými rámcami sme vykonávali pomocou funkcie *Layout*.

17 KATRENIAK, Martin: *Automatická transkripcia rukopisných historických textov na príklade vybraných kanonických vizitácií* [Diplomová práca]. Školiteľ O. Tomeček. Banská Bystrica : Univerzita Mateja Bela, 2022. 79 s.

Zo segmentácie textu sme vylúčili číslovanie v obrázkoch. Časová náročnosť na manuálnu segmentáciu a úpravu jednej dvojstrany bola približne 10 – 15 minút.

Po nasegmentovaní dokumentu sme pristúpili k *manuálnemu prepisu vybraných strán*. Na ciele našej práce sme si zvolili metódu transliterácie. Transkripcia a transliterácia sú jazykovedné pojmy, ktorých význam sa do veľkej miery prekrýva, niekedy sa dokonca nesprávne zamieňajú. Transkripcia v užšom zmysle znamená písomné vyjadrenie vyslovovaných alebo cudzím grafickým systémom zapísaných slov z hľadiska ich zvukovej stránky. Na druhej strane transliterácia alebo odborný/vedecký prepis sa primárne orientuje na vizuálnu podobu pôvodného textu (hoci často zachytáva aj fonetickú hodnotu pôvodnej grafémy). Má niekoľko definícií, v zásade však platí, že ide o „prevod z jednej grafickej sústavy do druhej, pri ktorom každému písmenu jedného grafického systému zodpovedá vždy písmeno druhého systému (rovnaké písmeno alebo spojenie písmen), takže je možný aj jednoduchý spätný prevod do jazyka originálu.“[18] Transliterácia poskytuje informáciu o tom, aké grafémy obsahuje pôvodný text, a umožňuje rekonštrukciu písanej podoby slova v pôvodnom texte. Výsledný prepis tlače preto môže byť atraktívny nielen pre historikov a literárnych vedcov, ale aj pre jazykovedcov či grafológov.

Naším hlavným cieľom bolo vytvoriť model, ktorý by umožňoval prepis všetkých fontov a jazykových mutácií textu súčasne. Tento postup sme zvolili aj z dôvodu menšieho rozsahu pôvodného dokumentu, ktorý slúžil na trénovanie modelu, tak aby trénovanie modelu vykazovalo čo najvyššiu mieru efektivity. V jednotlivých jazykoch bývajú zvyčajne vypracované zásady bežného a odborného prepisu z iných grafických sústav. Manuály na prácu so softvérom *Transkribus* odporúčajú prepis systémom písmeno za písmeno a v prípade stredovekých textov nepoužívanie modernej interpunkcie (v takýchto prípadoch je vhodnejšie interpunkciu úplne vynechať alebo používať špeciálne symboly).[19] Na účely transliterácie štvorjazyčnej historickej tlače v štyroch typoch písma sme si aj s ohľadom na vyššie uvedené vytvorili interné konvencie na prepis špeciálnych grafém znakmi UNICODE a určili, ktoré náhradné znaky sa použijú v prípade grafém, ktoré sústava UNICODE nepozná. Druhou zásadou, ktorú sme pri prepise uplatnili, bolo zachovanie tlačových chýb a všetkých nepresností, ktoré mohli vzniknúť priamo pri tlači, degradáciou papiera alebo nízkou kvalitou digitálnych snímok.

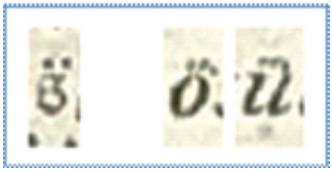
Historická tlač obsahuje množstvo znakov, ktoré sa v súčasnosti nepoužívajú. Najčastejšími boli ligatúry **æ** a **œ** v antikve a **ʧ** vo fraktúre. Všetky fonty písma obsahujú grafému **dlhé s** (ſ), ktorá v tom čase nahrádzala **okrúhle s** na začiatku a uprostred slabiky. Švabach obsahuje početne zastúpené grafémy **ċ**, **ė**, **ġ**, **ṅ**, **ṙ**, **ṫ**, **ż**, kurzíva grafémy **ö**, **ü** a fraktúra grafému **ostrého s** (ß). Všetky tieto grafémy majú v UNICODE svoj znak.

18 MISTRÍK, Jozef: Prepis z iného písma. In: *Encyklopédia jazykovedy*. Bratislava : Obzor, 1993, s. 343.

19 Transkribus Transcription Conventions. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-10-28]. Dostupné na: https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/#h-punctuation

Obrázok 58 Príklady ligatúr, dlhého s, ostrého s a iných grafém, pre ktoré má sústava UNICODE zodpovedajúce znaky. Zdroj: Orbis Pictus (1798).

Osobitným prípadom sú grafémy, ktoré štandardný UNICODE neobsahuje (obrázok 59). Vo švabachu sa vyskytuje graféma **s s dvomi bodkami**, ktorú sme v prepise nahradili významovým a fonetickým ekvivalentom š. Kurzíva vo veľkom počte obsahuje grafémy **u** a **o s tromi bodkami**. Systém UNICODE pozná znak ü, ale nie znak zodpovedajúci prepisu pre grafému **o s tromi bodkami**. Napriek našej snahe zachovať čo najvyššiu mieru transliterácie pri jednom z písmen nebolo možné zachovať jeho presný ekvivalent, preto sme z dôvodu zachovania jednotnosti pri prepise znakov zvolili náhradu oboch grafém významovo a foneticky obdobným znakmi **ő** a **ű**.



Obrázok 59 Príklady grafém fontu švabach a kurzíva, pre ktoré bolo potrebné vytvoriť jednotné pravidlo prepisu. Zdroj: Orbis Pictus (1798).

Fraktúra obsahuje grafémy **a s nadpísaným e**, **o s nadpísaným e** a grafému **u s nadpísaným e** (obrázok 60) s významom ä, ö, ü. Podobne ako grafémy kurzívy ani tieto nemajú v systéme UNICODE svoj ekvivalent. Na prepis týchto troch znakov sme vytvorili jednotné pravidlo a nadpísané e sme nahradili dvomi bodkami.

Obrázok 60 Príklady grafém fontu fraktúra, pre ktoré bolo potrebné vytvoriť jednotné pravidlo prepisu. Zdroj: Orbis Pictus (1798).

Po manuálnom prepise, ktorý by mal byť bezchybný, a kontrole prepísaných strán sa každá označí príznakom Ground Truth, čo znamená, že je pripravená na trénovanie. Takto označené strany by sa už nemali editovať ani inak meniť.

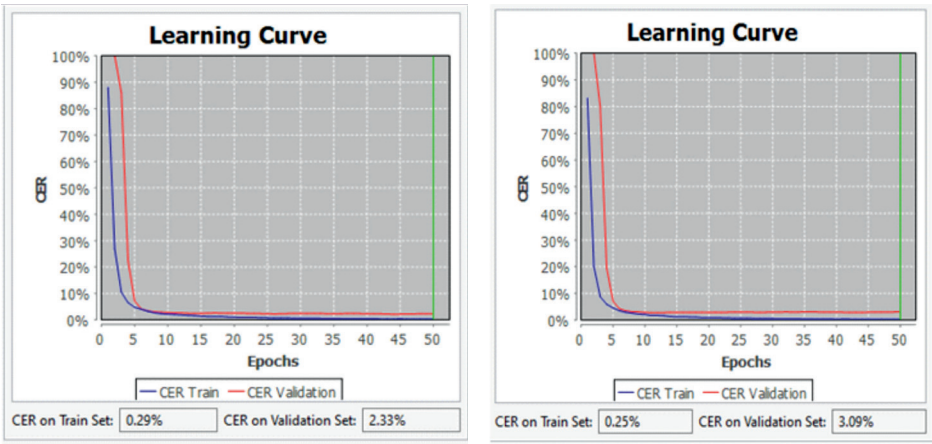
TVORBA A TRÉNOVANIE MODELOV

Odporúčaný počet transkribovaných slov v cvičnom súbore je 5 000 – 15 000 (približne 25 – 75 strán) v závislosti od toho, či ide o rukopisný alebo tlačený text. Strany sa do cvičného

a overovacieho súboru rozdeľujú spravidla v pomere 10 : 1. Model sa trénuje a overuje opakovane, štandardne je nastavených 50 cyklov (angl. Epochs). Ich počet môže používateľ podľa potreby zvyšovať alebo znižovať. Kvalita vytrénovaného modelu sa určuje pomocou hodnôt chybovosti čítania znakov CER (Character Error Rate) a slov WER (Word Error Rate), ktoré vyjadrujú priemerné percento znakov/slov chybne prepísaných softvérom.[20] Modely 1 – 8 sme vytrénovali pomocou technológie HTR+, Model 9 pomocou PyLaia.

POPIS MODELOV 1 A 2

Vzhľadom na to, že skúmaný dokument je tlač, nie rukopis a má relatívne malý rozsah (180 číslovaných strán a 93 naskenovaných dvojstrán obsahujúcich text), sme do modelov vybrali menší počet strán a pomer cvičného a overovacieho súboru sme rozdelili v pomere 5 : 1. Do Modelu 1 sme zaradili 10 strán, kde strany 1 – 8 (2 047 slov) boli v cvičnom a strany 9 – 10 (605 slov) v overovacom súbore. Keďže rozsah fontov na jednotlivých stranách sa menil, rozhodli sme sa súbežne pre ďalší model, ktorý sme označili ako Model 2. Zaradili sme doň tých istých desať strán s tým rozdielom, že v cvičnom súbore boli strany 1 – 6, 9 – 10 (1 878 slov) a v overovacom súbore strany 7 – 8 (773 slov).



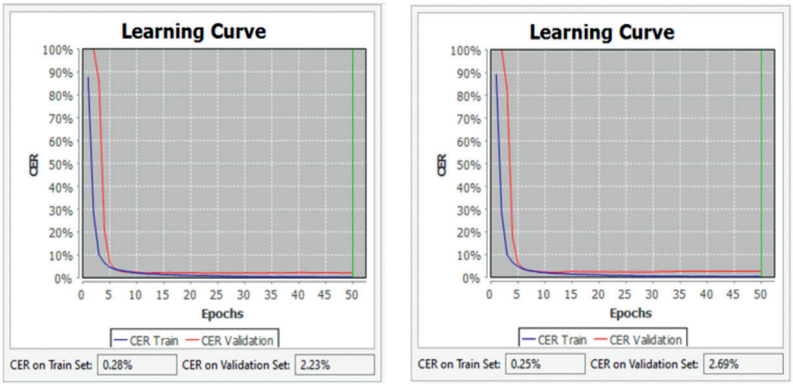
Obrázok 61 Porovnanie Modelu 1 (vľavo) a Modelu 2 (vpravo) na automatickú transkripciu diela Orbis Pictus (1798). Zdroj: *Transkribus.*

Výsledok trénovania CER 2,33 % a CER 3,09 % znamená, že v Modeli 1 bolo v cvičnom súbore bezchybne určených až 97,67 % znakov, v Modeli 2 o niečo menej – 96,91 %. Hodnoty CER automaticky prepísaných rukopisných strán do 10 % sa považujú za uspokojivé, v najlepších prípadoch sa pohybujú okolo 5 %. Vynikajúce výstupy modelov trénovaných na tlačených stranách dosahujú chybovosť približne 1 – 2 %. Z tohto dôvodu sme dosiahnuté výsledky považovali za neuspokojivé. Najväčšiu chybovosť sme pozorovali na stranách s vysokým výskytom textu v švabachu. Zároveň sme pri podrobnej analýze automatickej transkripcie v oboch modeloch zistili, že softvér upozornil aj na naše vlastné chyby pri manuálnej transliterácii.

20 MUEHLBERGER, Guenter et al. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: *Journal of Documentation*, vol. 75, no. 5, 2019, pp. 954-976.

POPIS MODELOV 3 A 4

Ako prvý spôsob zlepšenia modelov sme zvolili opravu chybne prepísaných slov. Takto upravené strany sme opätovne zaradili do trénovania. Model 3 s CER 2,23 % a Model 4 s CER 2,69 % na overovacom súbore však nepriniesli také zníženie chybovosti, aké sme očakávali.

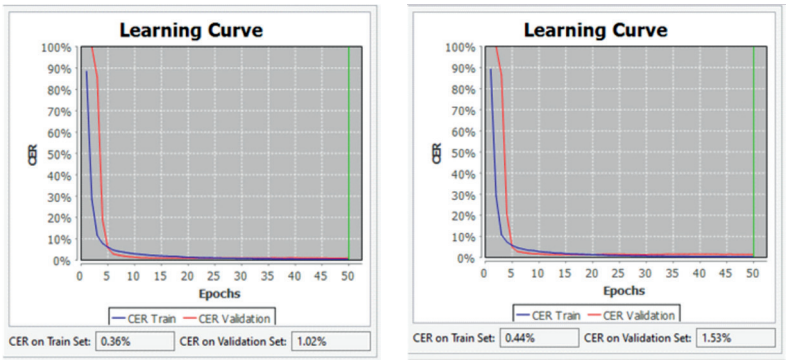


Obrázok 62 Porovnanie Modelu 3 (vľavo) a Modelu 4 (vpravo) pre automatickú transkripciu diela Orbis Pictus (1798). Zdroj: *Transkribus*.

POPIS MODELOV 5 A 6

Podľa princípu strojového učenia platí, že čím viac transkribovaných strán používateľ programu ponúkne, tým presnejšie výsledky získa. V prípade, že tlačený dokument obsahuje viacero jazykov, odporúča sa do vzorky Ground Truth zaradiť minimálne 5 000 prepísaných slov. Zo skúmaného dokumentu sme preto manuálne prepísali ďalších 5 strán, čím sme každú vzorku rozšírili o viac ako 2 000 slov na celkových 15 strán. Model 5 sme rozdelili na cvičný súbor s 12 stranami (1 – 8, 11 – 15) a overovací súbor s 2 stranami (9 – 10). V Modeli 6 sme do cvičného súboru zaradili strany 1 – 6, 9 – 15 a do cvičného súboru strany 7 – 8.

Dosiahnuté hodnoty CER 1,02 % a 1,53 % oboch modelov priniesli výsledky na úrovni najlepších modelov trénovaných na tlačených dokumentoch s jedným typom fontu.



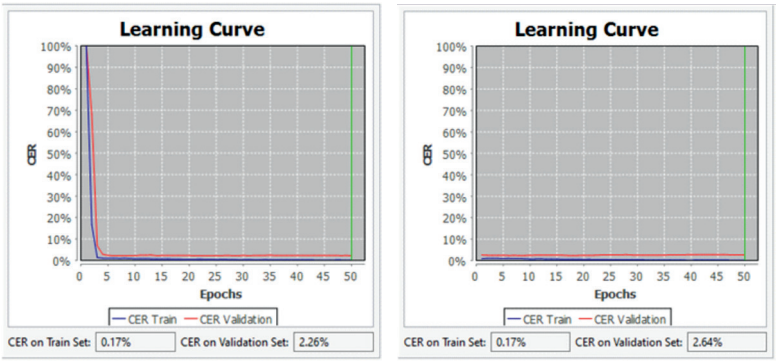
Obrázok 63 Porovnanie Modelu 5 (vľavo) a Modelu 6 (vpravo) na automatickú transkripciu diela Orbis Pictus (1798). Zdroj: *Transkribus*.

POPIS MODELOV 7 A 8

Ďalšou možnosťou zdokonaľovania modelu je použitie hotového modelu, ktorý sa označuje ako Base Model. Používateľ si môže vybrať z dvoch možností: vlastný vytrénovaný model alebo voľne dostupný model iného používateľa. Technológia OCR aplikovaná na moderné tlačené texty s využitím latinkového písma funguje s vysokou presnosťou a viac-menej sa považuje za vyriešený problém.[21] V prípade starých tlačí (a analogicky aj rukopisných dokumentov) však toto tvrdenie neplatí. Problém je najmä v neštandardnej typografii, vo vysokej variabilite dobového pravopisu, ktorá znemožňuje automatickú kontrolu pravopisu počas procesu OCR, a vo fyzickej degradácii papiera historickej tlače spôsobenej starnutím a používaním.[22] Model vytrénovaný na tlači so špecifickým fontom môže na tejto tlači fungovať s vynikajúcimi výsledkami, avšak jeho použitie na texte s iným fontom, ktorý sa ľudskému oku môže javiť ako podobný, nemusí priniesť očakávaný efekt.

Používateľská príručka k softvéru *Transkribus* upozorňuje, že pri výbere vhodného „cudzieho modelu" je najdôležitejším kritériom typ písma, či už rukopisu, alebo tlačového fontu. Pri práci s naším dokumentom sme dospeli k dvom zisteniam. Po prvé je dôležité zohľadniť typ fontu v kombinácii s jazykom textu. Vhodným príkladom je švabach, ktorý pri použití v textoch v slovanských jazykoch obsahuje špecifické grafémy a diakritické znamienka na samohláskach a spoluhláskach, ktoré sa napríklad v nemeckých textoch nenachádzajú. Po druhé je potrebné vziať do úvahy metódu prepisu, ktorá bola zvolená pri trénovaní Base Modelu a ktorú sme si vybrali my. Model trénovaný na texte prepísanom transkripciou nebude fungovať na texte, ktorý chceme transliterovať.

Model 7 a Model 8 sme trénovali technikou Base Modelu, pričom sme z vyššie uvedených dôvodov využili vlastné modely 1 a 2. Na základe výsledných hodnôt CER 2,26 % a 2,64 % na overovacom súbore môžeme konštatovať, že táto metóda priniesla len minimálne zlepšenie modelov a takmer žiadne zníženie chybovosti čítania znakov.



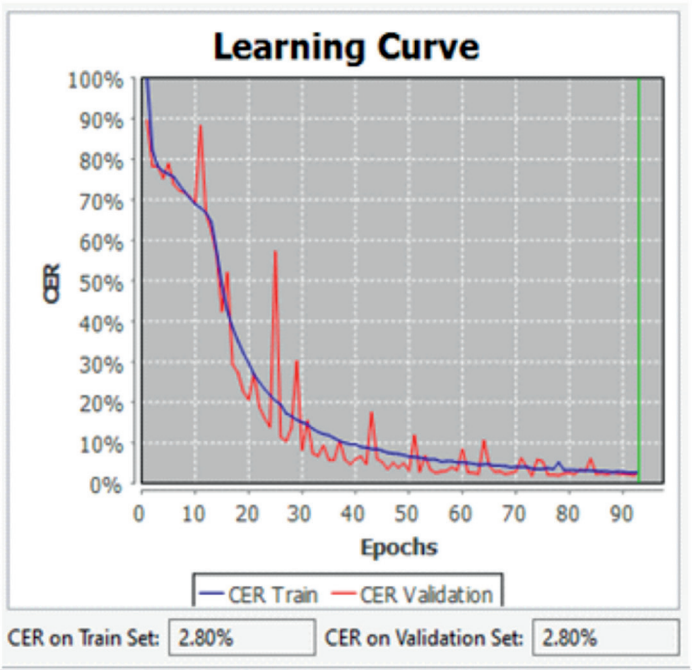
Obrázok 64 Porovnanie Modelu 7 (vľavo) a Modelu 8 (vpravo) pre automatickú transkripciu diela Orbis Pictus (1798). Zdroj: *Transkribus*.

21 DOERMANN, David – TOMBRE, Karl (eds.): *Handbook of Document Image Processing and Recognition* [online]. London : Springer Verlag, 2014 [cit. 2022-11-14]. 1055 p. Dostupné na: https://doi.org/10.1007/978-0-85729-859-1

22 SPRINGMANN, U. – LUDELING A.: OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. In: *arXiv. org* [online]. Ithaca : Cornell University, 2016 [cit. 2022-10-28]. 25 p. Dostupné na: https://doi.org/10.48550/arXiv.1608.02153

POPIS MODELU 9

Model 9 sme pripravili v zmysle metodických odporúčaní, preto sme do vzorky doplnili ďalšie strany na súhrnných 20, čím sa počet slov zvýšil na 5 791 v cvičnom súbore a 778 v overovacom súbore. Zachovali sme pomer cvičných a overovacích strán 10 : 1. Výhodou bolo, že na doplnených stranách sme v porovnaní s predchádzajúcimi zaznamenali vyšší výskyt ligatúr æ, ʧ a ß. Na porovnateľnosť s predošlými modelmi sme do overovacieho súboru vybrali stranu 8 z modelov 1, 3, 5 a 7 a stranu 10 z modelov 2, 4, 6, a 8. Zvýšením počtu slov nad odporúčanú hranicu a dodržaním pomeru cvičných a overovacích strán sme dosiahli výsledky s hodnotami CER 2,80 % na cvičnom a 2,80 % na overovacom súbore.



Obrázok 65 Model 9 na automatickú transkripciu diela *Orbis Pictus* (1798). Zdroj: *Transkribus.*

V čase trénovania Modelu 9 (november 2022) prešla platforma *Transkribus* zásadnou zmenou. Z dôvodu časovej náročnosti na udržiavanie proprietárneho softvéru a zastaranosti kódu pristúpili vývojári k odstaveniu HTR+ technológie. Na trénovanie nových modelov aktuálne možno použiť už len technológiu PyLaia. Príprava tré-novacej vzorky prebieha rovnako ako pri modeloch HTR+. Aj v prípade technológie PyLaia sa odporúča začať trénovať s 5 000 až 15 000 slovami transkribovaného textu v závislosti od toho, či ide o tlačený alebo rukopisný dokument. Použitie Base Modelu na vytrénovanie nových modelov znižuje množstvo slov potrebných na trénovanie a ako Base Model je možné použiť výlučne model vytrénovaný technológiou PyLaia. Hotové HTR+ modely sa na PyLaia technológiu dajú pretrénovať prostredníctvom platformy *Transkribus* Lite.

PyLaia ponúka viac funkcií najmä pre skúsených tvorcov modelov s možnosťami manuálneho nastavenia viacerých parametrov. Jedným z nich je funkcia *Deslant*, ktorá má za úlohu vyrovnávať kurzívny typ písma charakteristický predovšetkým pre rukopisné texty. Ak sa však kurzívne písmo vyskytuje v tlačenom dokumente, odporúča sa funkciu nepoužiť, pretože efekt môže byť opačný.[23] Keďže náš dokument obsahuje štvrtinu textu vo fonte kurzíva, túto funkciu sme z prednastavenia odstránili.

Na základe výsledkov Modelu 9, ktorý z deviatich trénovaných modelov vyšiel v parametri WER ako najhorší a v parametri CER ako druhý najhorší, je otázne, či skutočne platí, čo deklarujú autori platformy, že obe technológie pracujú podobne a výsledky vyjadrené hodnotou CER sú zvyčajne rovnaké. Pretrénovanie Modelov 1 – 8 a porovnanie výsledkov získaných technológiou HTR+ a PyLaia môže byť námetom na ďalší výskum v rámci tohto projektu.

VYHODNOTENIE MODELOV

Keďže do modelov 1, 3, 5, 7 a 2, 4, 6, 8 sme zaradili tie isté overovacie strany, výsledky na základe hodnôt WER a CER sú jednoznačne porovnateľné.

Tabuľka 11 Prehľad Character Error Rate (CER) a Word Error Rate (WER) v cvičných a overovacích súboroch Modelov 1 – 9.

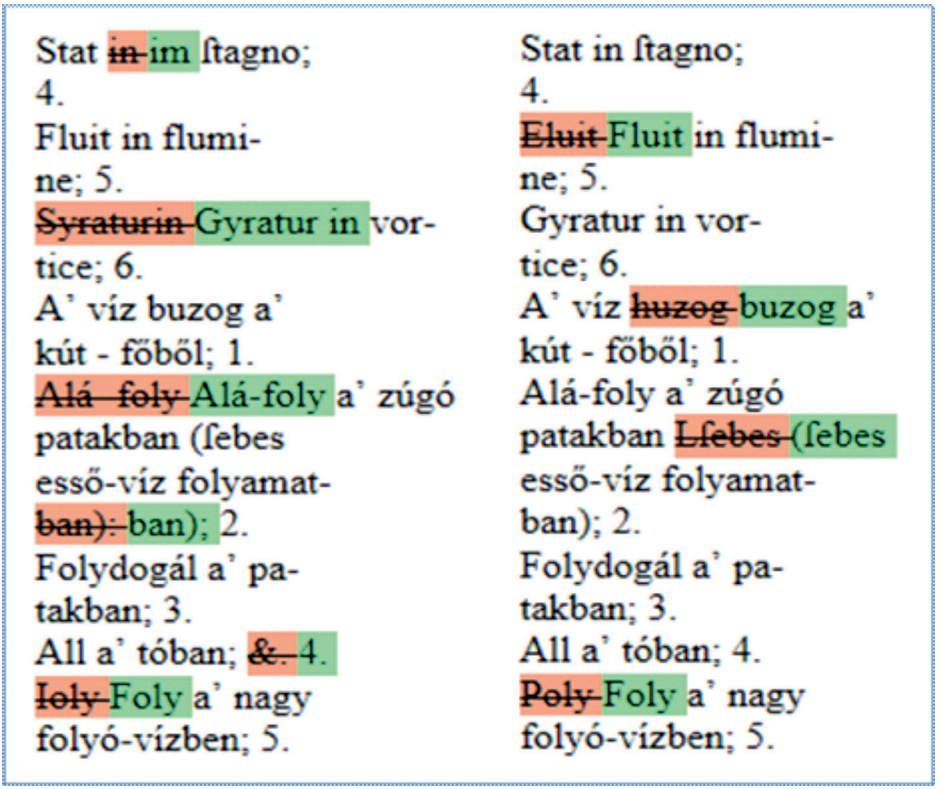
Model	ID modelu	Technológia	Cvičný súbor			Overovací súbor				
			počet riadkov	počet slov	CER	počet riadkov	počet slov	CER	WER 1	WER 2
Model 1	43995	CITlab HTR+	653	2 047	0,29 %	202	605	2,33 %	6,42 %	8,56 %
Model 2	44001	CITlab HTR+	601	1 878	0,25 %	254	773	3,09 %	6,65 %	9,74 %
Model 3	44136	CITlab HTR+	653	2 047	0,28 %	202	605	2,23 %	7,74 %	7,31 %
Model 4	44117	CITlab HTR+	601	1878	0,25 %	254	773	2,69 %	6,65 %	9,00 %
Model 5	44220	CITlab HTR+	1 318	4 107	0,36 %	202	603	1,02 %	2,43 %	3,34 %
Model 6	44218	CITlab HTR+	1 266	3 938	0,44 %	254	772	1,53 %	5,08 %	3,72 %
Model 7	44247	CITlab HTR+	653	2 049	0,17 %	202	603	2,26 %	4,87 %	8,35 %
Model 8	44251	CITlab HTR+	601	1 880	0,17 %	254	772	2,64 %	6,85 %	9,14 %
Model 9	46060	PyLaia	1 826	5 791	2,80 %	248	778	2,80 %	12,63 %	7,30 %

ANALÝZA CHÝB

Zaradenie tých istých strán do overovacích súborov jednotlivých modelov (s výnimkou Modelu 9) sa ukázalo ako dobré riešenie nielen z hľadiska porovnateľnosti hodnôt CER a WER. Zároveň sme získali aj kvalitné podklady na podrobnú analýzu chybovosti, ktorá priniesla niekoľko zaujímavých zistení.

Najprekvapujúcejší bol vznik nových chýb v Modeloch 2 a 4 v porovnaní s Modelmi 1 a 3. Odstránením chýb manuálnej transliterácie sme očakávali lepšie výsledky vytrénovaných modelov. Sedem nových chýb v Modeli 3 a až osemnásť v Modeli 4 však ich úspešnosť výrazne znížili. V Modeli 3 softvér správne prečítal pätnásť slov, v ktorých urobil chybu v Modeli 1 a v Modeli 4 sedemnásť slov v porovnaní s Modelom 2.

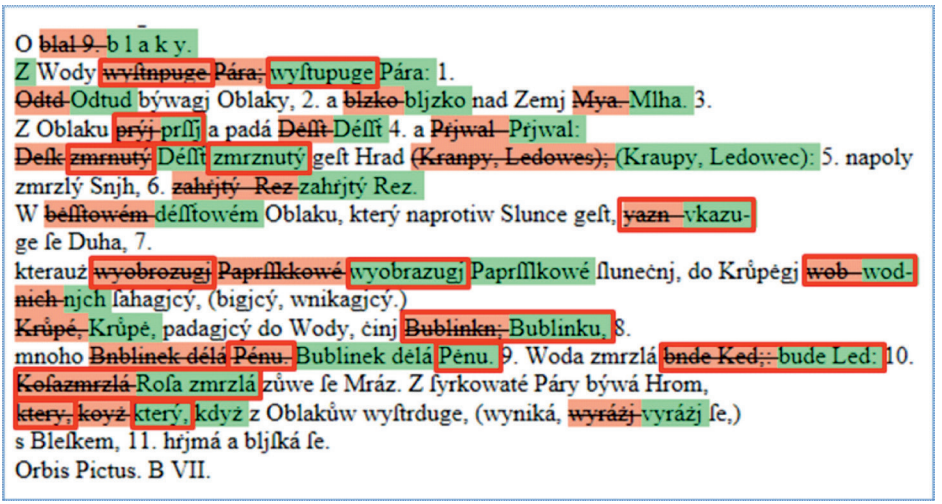
23 How to Train PyLaia-Models in Transkribus. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-11-09]. Dostupné na: https://readcoop.eu/transkribus/howto/how-to-train-pylaia--models-in-transkribus/



Obrázok 66 Porovnanie nových a odstránených chýb na fragmente textu (Model 2 vľavo, Model 4 vpravo). Zdroj: *Transkribus.*

V každom ďalšom modeli sa úspešnosť prepisu niektorých fontov zvyšovala, vznikali však aj nové chyby: v Modeli 5 dve chyby v porovnaní s modelmi 1 a 3, v Modeli 6 jedenásť chýb v porovnaní s modelmi 2 a 4, v Modeli 7 dve k modelom 1, 3 a 5, v Modeli 8 trinásť k modelom 2, 4, a 6. Chyby v modeloch neboli úplne totožné. Napriek aplikácii rôznych metód pri vylepšovaní modelov sa niektoré časti textu ukázali ako vysoko problematické a chyby sa v nich opakovali naprieč všetkými modelmi. Najčastejšie sme zaznamenali nesprávny prepis interpunkcie a špecifických grafém vo fonte švabach a kurzíva, v kurzíve chybný prepis apostrofu a vo všetkých fontoch zámenu za iné písmená, najmä a/u, ſ /l, ſ/f. V ojedinelých prípadoch sa vyskytol problém s doplnením písmen, ktoré sa v originálnom texte nenachádzali, napr. Extra namiesto Ex. Domnievame sa, že výsledok opakovanej aplikácie toho istého modelu na rovnakej vzorke overovacích strán nebude vždy totožný.

Pri trénovaní modelu pomocou technológie PyLaia sa objavili problémy so správnym čítaním niektorých grafém, napr. grafému u si softvér najmä vo švabachu zamieňal s grafémou n, vo fraktúre N s R, niektoré grafémy z prepisu úplne vynechal. Výskyt viacerých chýb je porovnateľný s inými modelmi s vyššou chybovosťou.


O blal 9. b l a k y.
Z Wody wyſtnpuge Pára, wyſtupuge Pára: 1.
Odtd Odtud bywagj Oblaky, 2. a blzko bljzko nad Zemj Mya Mlha. 3.
Z Oblaku prýj prſlj a padá Déſſt Déſſt 4. a Přjwal Přjwal:
Deſk zmrnutý Déſſt zmrznutý geſt Hrad (Kranpy, Ledowes), (Kraupy, Ledowec): 5. napoly
zmrzlý Snjh, 6. zahrjtý Rez zahrjtý Rez.
W beſſtowém déſſtowém Oblaku, který naprotiw Slunce geſt, yazn vkazu-
ge ſe Duha, 7.
kterauż wyobrozugj Paprſſkkowé wyobrazugj Paprſſlkowé ſlunečnj, do Krůpégj wob wod-
nich njch ſahagjcý, (bigjcý, wnikagjcý.)
Krůpé, Krůpě, padagjcý do Wody, činj Bublinku, Bublinku, 8.
mnoho Bnblinek dělá Pénu. Bublinek dělá Pěnu. 9. Woda zmrzlá bnde Ked, bude Led: 10.
Koſazmrzlá Roſa zmrzlá zůwe ſe Mráz. Z ſyrkowaté Páry bývá Hrom,
ktery, koyż který, když z Oblaků wyſtrduge, (wyniká, wyrážj wyrážj ſe,)
s Bleſkem, 11. hřjmá a bljſká ſe.
Orbis Pictus. B VII.

Obrázok 67 Nové chyby v Modeli 9 (v červenom orámovaní), ktoré sa v iných modeloch nevyskytovali. Zdroj: *Transkribus.*

Chyby vo vytrénovaných modeloch sme rozdelili do troch kategórií: chybný prepis grafémy, chybný prepis interpunkcie, chybná identifikácia medzery.

Tabuľka 12 Prehľad najčastejšie sa opakujúcich chýb pri prepise grafémy.

Nesprávne	Správne
Flüfte	Klüfte
Acerfeld	Ackerfeld
Heudel-	Heydel-
dte Erdſchwámme,	die Erdſchwämme,
fæ-	fœ-
fctus	Fœtus
Érűznemek	Értz-nemek
füvekkel	fűvekkel
Jahody	Gahody
twrdė	twrdé
Leċergetky	Ċeċergetky
zžirá	zžjrá
a	a'

Analýza chýb vo všetkých trénovaných modeloch ukázala, že nejde o závažné chyby. V prípade textu napísaného v latinskom, maďarskom a nemeckom jazyku však sťažujú jeho čítanie. Z hľadiska transliterácie však chyby, ktoré súvisia s nesprávnym prepisom grafémy, považujeme za závažné. Najčastejšie sa vyskytovali vo fonte švabach (najmä č namiesto ċ, é namiesto ė, J namiesto G, i namiesto j). Tieto grafémy sú v texte zastúpené vo veľkom počte, pretože bodky nad písmenami nahrádzajú diakritické znamienka mäkčeň a dĺžeň. V kurzíve sa nesprávny prepis grafémy vyskytoval najmä v prípade zámeny ü/ű a ö/ő. Tento problém možno čiastočne odstrániť zdokonaľovaním modelu na väčšom počte strán, vďaka čomu sa softvér naučí tieto znaky správne rozpoznať.

Tabuľka 13 Prehľad najčastejšie sa opakujúcich chýb pri čítaní interpunkcie.

Nesprávne	Správne
(Audolj);	(Audolj):
fruges;	fruges,
olera-	olera.

V niektorých kapitolách tejto publikácie sa autori (Nagy, Tomeček) zmieňujú, že program *Transkribus* má problém s čítaním číslic. V modeloch vytrénovaných na našej historickej tlači pri použití HTR+ technológie sa tento problém nepotvrdil. Zastúpenie číslic je pomerne vysoké, pretože okrem paginácie sú významnou súčasťou textovej časti, ktorá pomocou číselných odkazov prepája vysvetľované pojmy s obrázkami zobrazujúcimi tému kapitoly. Chyby sa výnimočne vyskytovali pri čísliciach 1, 2, 5, 8 a 9 (napr. 10 namiesto 19, 1 namiesto 21, 15 namiesto 18). Pri použití technológie PyLaia sa chyby vyskytovali pri čísliciach 3 a 7 (5 namiesto 7, 3 namiesto 7), dokonca pri nesprávnom prepise písmena na číslicu (3 namiesto z, 9 namiesto y).

Tabuľka 14 Prehľad najčastejšie sa opakujúcich chýb pri identifikácii medzery.

Nesprávne	Správne
dieStei-	die Stei-
ſcuntur :	ſcuntur:
O bla k y.	O b l a k y.

V originálnej tlači sa vo zvýšenej miere vyskytujú medzery medzi slovom a nasledujúcim interpunkčným znamienkom (čiarka, dvojbodka, bodkočiarka). Paginácia je umiestnená v zátvorkách, ktoré sú od čísla strany oddelené medzerou. Medzery tlačiar využil aj na zvýraznenie nadpisu v českom jazyku. Pri transkripcii sme sa snažili zachovať medzery výlučne pri paginácii a nadpise kapitol, v prípade interpunkcie sme ich pri prepise ignorovali. Pri analýze chybovosti automatického prepisu sme zistili, že s odstránením

medzery pri interpunkcii sa program dokázal vyrovnať a v overovacích súboroch sa súvisiace chyby vyskytovali len ojedinele. Na porozumenie textu nemali závažný vplyv, nijako neznižovali kvalitu vytrénovaného modelu a v podstate išlo len o nami stanovenú konvenciu transkripcie. Za závažnejšie chyby prepisu možno považovať absenciu priznaných medzier v texte oddeľujúcich slová v texte, za menej závažné chýbajúce medzery medzi písmenami v názvoch kapitol.

ZÁVER

Výsledky ukazujú, že pri modeloch 5 a 6 sa ako najvhodnejšia metóda na zdokonalenie spoločného modelu pre štyri fonty javilo zvýšenie počtu strán, a teda aj častejší výskyt jednotlivých grafém v cvičnom súbore. V tejto fáze sa tým potvrdila téza, že čím viac cvičných dát programu poskytneme, tým kvalitnejšie výsledky získame. Model 9 však prekvapil a svedčí o tom, že ani dodržanie odporúčaného počtu strán/slov a pomeru cvičnej a overovacej vzorky nemusí nevyhnutne viesť k vytrénovaniu vynikajúceho modelu. Domnievame sa, že tieto pravidlá sú lepšie uplatniteľné na dokumentoch s homogénnejším písmom.

Tlačený dokument, ktorý obsahuje viac fontov, možno prirovnať k rukopisným textom s viacerými rukopismi. Pri trénovaní spoločných modelov na takomto texte záleží na výbere strán do cvičného a overovacieho súboru. Výsledky totiž významne ovplyvní to, aké množstvo textu v problematickom písme (príp. písmach) sa v nich nachádza a v akom pomere.

Ako najproblematickejšie písmo sa ukázal švabach, ktorý zvyšoval chybovosť čítania znakov vo všetkých spoločných modeloch. Dokazujú to aj hodnoty namerané na samostatnom trénovaní fontov, kde bola miera chybovosti fontu antikva 2,64 % na overovacom súbore, fontu kurzíva 2,11 %, fontu fraktúra 1,23 % a fontu švabach až 4,78 %.

Výsledky Modelu 9 trénovaného pomocou technológie PyLaia nie sú z dôvodu rozdielnosti trénovacích nástrojov s prechádzajúcimi modelmi úplne porovnateľné. Napriek väčšiemu počtu strán zahrnutých do cvičnej a overovacej vzorky sme očakávali lepšie výsledky. V porovnaní s HTR+ modelmi sa v modeli Pylaia objavil problém so správnym čítaním aj bežných grafém.

Výberu Base Modelu od iného používateľa musí predchádzať dôkladná analýza postupov a metód, ktoré boli na jeho vytrénovanie použité. Ak základný model nemá špecifické znaky pre fonty, ktoré skúmaný dokument obsahuje (v našom prípade pre český švabach), jeho využitie na automatickú transkripciu je neopodstatnené.

Automatický prepis dokumentu si (aj pri najlepších modeloch) vyžaduje následnú opravu chýb v transkribovanom texte. Potvrdilo sa, že z hľadiska čítania a porozumenia textu vo väčšine prípadov nejde o závažné chyby. Chýbajúcu diakritiku, interpunkciu či nesprávnu grafému možno opraviť spôsobom bežnej jazykovej korektúry. Nevýhodou je, že v týchto prípadoch sa nedá ako pomôcka využiť automatická kontrola pravopisu v textovom editore, a to z dôvodu dobového stavu jazyka (jazykov) dokumentu a transliterácie ako metódy prepisu.

Využiteľnosť a efektivita vytrénovaného modelu je tým vyššia, čím viac strán sa ním bude automaticky prepisovať. S vyšším počtom slov sa zvýšil aj výskyt špeciálnych znakov a výrazné zníženie chybného čítania ligatúry tʒ sme zaznamenali v Modeloch 5, 7 a 9. Pri tlačiach s menším počtom strán, resp. slov, kde je potrebné manuálne prepísať minimálne 25 strán, sa využiteľnosť modelu na automatický prepis zvyšku dokumentu v pomere k času potrebnému na formálnu prípravu dokumentu (segmentácia), vzorky Ground Truth (manuálna transkripcia) a vytrénovanie samotného modelu významne znižuje.

Domnievame sa, že opakované trénovanie modelu pri zachovaní tých istých podmienok (tie isté strany v cvičnom i overovacom súbore, rovnaký počet opakovaní trénovania) bude vykazovať rôzne výsledky. Overenie tejto teórie môže byť tiež námetom na ďalší výskum.

Program *Transkribus* môže byť vo všeobecnosti využiteľný v oblasti občianskej vedy. „Občianska veda zapája laickú verejnosť do riešenia vedeckých projektov. Spôsoby zapojenia občanov aj formy vedeckých projektov sú rôzne, často sprostredkované modernými technológiami, vďaka čomu sa občianskym vedcom/vedkyňou môže stať ktokoľvek – žiaci a študenti, pedagógovia, amatérski záujemcovia o vedu, ľudia pracujúci v iných odvetviach či seniori. Občania ako dobrovoľníci sa najčastejšie aktívne podieľajú na zbere materiálu, údajov alebo pozorovaniach v teréne, prípadne na spracovaní dát pomocou mobilných aplikácií.“[24] V tomto prípade sa občianski dobrovoľníci môžu podieľať de facto na všetkých fázach spracovania dokumentu programom *Transkribus* bez toho, aby museli byť odborníci na ktorúkoľvek oblasť historických vied: na skenovaní tlačených dokumentov (digitalizácia), na formálnej úprave dokumentu (automatická a manuálna segmentácia textu), jazykovo zdatní aj na manuálnom prepise vzorky Ground Truth či na jazykovej redakcii už transkribovaného textu. Takáto spolupráca má množstvo benefitov, okrem pomoci pri digitalizácii archívnych fondov, ktorá má na Slovensku stále veľké rezervy[25], aj možnosť vzdelávať sa a v praxi získavať nové zručnosti. Je to jedna z ciest, ako budovať dôveru medzi vedeckou komunitou a verejnosťou.

ZÁMERY V RÁMCI POKRAČOVANIA VÝSKUMU

Zmena technológie na platforme *Transkribus* si vyžiada pretrénovanie najlepších modelov, ktoré sme pôvodne trénovali pomocou technológie HTR+. Tento moment využijeme ako príležitosť na vylepšenie metódy transliterácie prostredníctvom možností, ktoré ponúka kombinovaný UNICODE[26]. K dispozícii má špeciálne znaky ako nadpísané e nad samohláskami, spoluhlásky s dvomi/tromi bodkami a i. Odstránime tak znaky štandardného súboru, ktoré nahrádzali pôvodné grafémy, a tým zvýšime autenticitu ortografického prepisu do maximálnej možnej miery.

Po automatickej transkripcii celého textu bude potrebné do výsledného dokumentu doplniť metadáta, ktoré sú nevyhnutnou súčasťou (nielen) zdigitalizovaných dokumentov dostupných v prostredí internetu a sú užitočné predovšetkým na vyhľadávanie vo veľkom

24 Občianska veda. In: *Otvorená veda* [online]. Bratislava : Centrum vedecko-technických informácií SR [cit. 2022-11-03]. Dostupné na: https://otvorenaveda.cvtisr.sk/obcianska-veda/

25 NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 53 – 67.

26 Online Unicode Tools. Dostupné na: https://onlineunicodetools.com/add-combining-characters

množstve dostupných informácií. Platforma *Transkribus* ponúka aj rozhranie na tvorbu textových metadát, tzv. tagovanie, ktoré umožňuje podrobnejší popis prepísaného textu. V ďalšej fáze vývoja a aplikovania modelu chceme otestovať funkciu tagovania transkribovaného dokumentu prostredníctvom preddefinovaných a/alebo vlastných kategórií značiek, ktorými je možné označovať dôležité slová, osoby, geografické miesta, frázy, skratky, typografické ozdoby, obrázky, ale aj nečitateľné grafémy či výrazy. Takto označené časti textu zvyšujú jeho prehľadnosť pre koncového používateľa.

Existuje viacero spôsobov a postupov, ako vytvoriť model na automatický prepis textu. Z hľadiska ďalšej využiteľnosti hotových modelov je dôležité, aby bol tento proces transparentne popísaný. Na tento účel ponúka program *Transkribus* špeciálnu funkciu *redakčné vyhlásenie* (angl. Editorial Declaration). Obsahuje súbor preddefinovaných funkcií s možnosťou tvorby vlastných popisov. V redakčnom vyhlásení sa uvádza aj zoznam znakov UNICODE použitých na prepis špeciálnych znakov, ktoré dokument obsahuje. Tento krok je nevyhnutný najmä v prípade transliterácie historických rukopisov a historických tlačí, pretože ďalším používateľom uľahčuje interpretáciu vytrénovaného modelu, rozhodovanie o jeho použití na prepis iných dokumentov, prípadne na opravu chýb už prepísaného dokumentu.

Zavŕšením prác na automatickej transkripcii historickej tlače pomocou programu *Transkribus* by malo byť jej sprístupnenie odbornej a laickej verejnosti.

Vzhľadom na technologické zmeny na platforme bude potrebné zvážiť ďalší postup týkajúci sa trénovania a využívania funkčných modelov. Model vytrénovaný pomocou HTR+ technológie nemusí byť po pretrénovaní v PyLaia rovnako úspešný. Keďže Model 9 (PyLaia) vykazoval pomerne vysokú chybovosť, bude predmetom nášho výskumu overenie úspešnosti najlepšieho HTR+ modelu po pretrénovaní technológiou PyLaia. Rovnako budeme postupovať v prípade samostatného modelu pre font švabach.

V historickom fonde univerzitnej knižnice sa nachádza aj neskoršie vydanie učebnice Orbis Pictus z tlačiarne Š. P. Webera z roku 1820. Dokumenty sú obsahovo totožné, rozdelené do rovnakého počtu kapitol, sú v nich použité tie isté drevorezy, text je napísaný v štyroch jazykoch a vytlačený v štyroch fontoch. Odlišná je však kvalita tlače, preto sa domnievame, že najlepší model bude vhodné otestovať práve na tomto vydaní a porovnať mieru chybovosti prepisu. Model bude tiež možné využiť na prepis historickej tlače z 18. storočia z dielne iného prešporského kníhtlačiara Jána Pavla Royera *Adparatvs ad Historiam Hvngariae*[27] z roku 1735.

Z grafologickej a obsahovej stránky by bolo zaujímavé porovnať vydanie Orbis Pictus z rokov 1798 a 1820 s prvým vydaním na Slovensku, ktoré vyšlo v Brewerovej kníhtlačiarni v Levoči v roku 1685. Ako sme už uviedli v inej časti tejto kapitoly, obsahoval prvý český preklad tohto významného diela. Pomerne dobre je zdokumentované pozadie vzniku druhého vydania z lisu tej istej kníhtlačiarne v roku 1728.[28]

27 BEL, Matej: *Adparatvs ad Historiam Hvngariae, sive collectio miscella, Monumentorum ineditorum partim, partim editorum, sed fugientium. Conquisiut, in Decades partitus est, & Praefationibus, atque Notis illustrauit, Matthias Bel*. Posonii : Typis Joannis Paulli Royer, A. MDCCXXXV [1735].

28 ČUMA, Andrej. Jana Amosa Komenského Orbis Pictus stále aktuálny. In: *Pedagogika: časopis pro pedagogické vědy*, roč. 41, č. 5 – 6, 1991, s. 603 – 609.

Poďakovanie

Autorky ďakujú doc. PhDr. Petrovi Voitovi, CSc., literárnemu historikovi, archivárovi, knihovníkovi a odborníkovi na historické knižné fondy za ochotu a identifikáciu fontov fraktúra a švabach. Mgr. Adriane Matejkovej, PhD., za nezištnú pomoc a čas pri hľadaní informácií o skúmanej tlači v historickom katalógu lyceálnej knižnice v Banskej Štiavnici. Mgr. Zuzane Denkovej, PhD., riaditeľke Slovenského banského múzea v Banskej Štiavnici za nasmerovanie a podporu.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

BAKOŠ, Ľudovít: Didaktické aspekty Komenského „Orbis Pictus“. In: *Acta Facultatis Philosophicae Univeritatis Šafarikanae Prešovensis, Pedagogika II/1971*. red. F. Karšai, Košice : Východoslovenské vydavateľstvo v Košiciach, 1971, s. 143 – 153.

BOKROSOVÁ, Katarína: Zbierka Komenián zo 17. až 19. storočia v študovni historických tlačí Slovenskej pedagogickej knižnici v Bratislave. In: *Bibliotheca Antiqua* [online]. Olomouc : Vědecká knihovna v Olomouci, 2014, s. 30 – 36 [cit. 2022-11-14]. Dostupné na: https://vmdelta.vkol.cz/data/soubory/hf/bibliotheca-antiqua-14/03bokrosova%20.pdf

CESNAKOVÁ-MICHALCOVÁ, Milena: Šimon Peter Weber: portrét bratislavského kníhtlačiara a dramatika. In: *Slovenské divadlo : revue dramatických umení*, roč. 24, č. 4, 1976, s. 571 – 587.

COMENII, Ioann. Amos: *Orbis Pictus, in hungaricum, germanicum et slavicum translatus, et hic ibive emendatus*. Posonii : Sumtibus & Typis Simonis Petri Weber, 1798. Zdigitalizované dostupné na: https://www.slovakiana.sk

COMENII, Joh. Amos: *Orbis Sensualium Pictus Quadrilinguis*. Leutschoviae : Typis Samuelis Brewer, 1685. Zdigitalizované dostupné na: https://www.slovakiana.sk

COMMENII, Joh. Amos: *Orbis Sensualium Pictus*. Noribergae : Typis & Sumptibus Michaelis Endteri, 1658. Zdigitalizované dostupné na: http://diglib.hab.de/drucke/47-7-eth-as/start.htm

COMENIUS, Iohannes Amos: Orbis Sensualium Pictus. nem. preklad Siegmund von Birken, ilustr. Paulus Creutzberger. In: *Bibliotheca Augustana* [online]. [cit. 2022-11-20]. Dostupné na:

http://www.hs-augsburg.de/~harsch/Chronologia/Lspost17/Comenius/com_o000.html

ČUMA, Andrej: Jana Amosa Komenského Orbis Pictus stále aktuálny. In: *Pedagogika : časopis pro pedagogické vědy*, roč. 41, č. 5 – 6, 1991, s. 603 – 609.

DECKER, Viliam: *Dejiny ručnej výroby papiera na Slovensku.* Martin : Matica slovenská, 1982. 223 s.

DOERMANN, David – TOMBRE, Karl (eds.): *Handbook of Document Image Processing and Recognition* [online]. London : Springer Verlag, 2014 [cit. 2022-11-14]. 1055 p. Dostupné na: https://doi.org/10.1007/978-0-85729-859-1

HOMOLOVÁ, Ľudmila: *Sprievodca historickým knižničným fondom Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici.* Banská Bystrica : Univerzita Mateja Bela v Banskej Bystrici, 2012. 48 s.

How to Train PyLaia-Models in Transkribus. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE, 2021 [cit. 2022-11-09]. Dostupné na: https://readcoop.eu/Transkribus/howto/how-to-train-pylaia-models-in-Transkribus/

KATRENIAK, Martin: *Automatická transkripcia rukopisných historických textov na príklade vybraných kanonických vizitácií* [Diplomová práca]. Školiteľ O. Tomeček, Banská Bystrica : Univerzita Mateja Bela, 2022. 79 s.

KATUŠČÁK, Dušan: Digital Humanities a automatická transkripcia rukopisných textov. In: *ITLib: informačné technológie a knižnice*, roč. 24, č. 1, 2020, s. 6 – 16.

KOWALSKÁ, Eva: Učebnice pre štátne ľudové školy na Slovensku koncom 18. storočia. In: *Kniha ´90. Zborník o problémoch a dejinách knižnej kultúry na Slovensku*. zost. M. Domová - R. Brož, Martin : Matica slovenská, 1990, s. 63 – 77.

MÁTEJ, Jozef: Pedagogické dedičstvo J. A. Komenského na Slovensku. In: *Acta Facultatis Philosophicae Universitatis Šafarikanae Prešovensis, Pedagogika II/1971.* red. F. Karšai, Košice : Východoslovenské vydavateľstvo v Košiciach, 1971, s. 27 – 45.

MISTRÍK, Jozef: Prepis z iného písma. In: *Encyklopédia jazykovedy.* Bratislava : Obzor, 1993. 513 s.

MUEHLBERGER, Guenter et al.: Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: *Journal of Documentation,* vol. 75, no. 5, 2019, pp. 954-976.

NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*, roč. 51, č. 2, 2021, s. 53 – 67.

Občianska veda. In: *Otvorená veda* [online]. Bratislava : Centrum vedecko-technických informácií SR [cit. 2022-11-03]. Dostupné na: https://otvorenaveda.cvtisr.sk/obcianska-veda/

PAVÚK, Marián: Kachelmannovci. In: *Haló Vyhne* [online], č. 3, 2023, s. [1 – 2] [cit. 2022-10-24]. Dostupné na: https://www.vyhne.sk/kachelmannovci/mid/66591/.html#m_66591

PŠENÁK, Jozef: Jan Amos Komenský, tvorca školských divadelných hier. In: *Paedagogica. Zborník Filozofickej fakulty Univerzity Komenského*, roč. 17. red. Š. Švec, Bratislava : Univerzita Komenského v Bratislave, 2005, s. 127 – 138.

SPRINGMANN, U. – LUDELING A.: OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. In: *arXiv.org* [online]. Ithaca : Cornell University, 2016 [cit. 2022-10-28]. 25 p. Dostupné na: https://doi.org/10.48550/arXiv.1608.02153

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

ŠTURDÍKOVÁ-HUDÁKOVÁ, Marta: Bratislavskí tlačiari 18. storočia a ich vyzdobené tlače. In: *Kniha ´82. Zborník pre problémy a dejiny knižnej kultúry na Slovensku.* zost. J. Valach, Martin : Matica slovenská, 1983, s. 103 – 112.

Transkribus Transcription Conventions. In: *READ-COOP* [online]. Innsbruck : READ-COOP SCE [cit. 2022-10-28]. Dostupné na: https://readcoop.eu/Transkribus/howto/Transkribus-transcription-conventions/#h-punctuation

Tabuľka 15 Prehľad najlepších modelov v projekte SKRIPTOR (2020 – 2022)[29].

Vlastník modelu (owner)	Názov dokumentu v TRB	Rok(y) vznikú/ pokrytie	Počet str.: dok	Jazyk/ jazyky	Typ písma RKP/PRT	R uky	Názov modelu	ID modelu	Počet slov TR AIN	Počet riadkov TR AIN	CER TRAIN	CER VALID	Stroj
MALINIAK	Postila Izáka Abrahamidesa	1600/1601	722	slo/cze; lat	novogotické humanistické písmo (RKP	1	Abraham model 3	112045 5	11434	1817	1,56%	7,99%	HTR+
NAGY	Csákóso v katalóg korešpondencie Kohárycov	1944/1945	4140	lat/hun/ger	moderná kurzíva	1	SKRIPTOR_Csákos_Koháry_3	45232	16307	2800	0.98%	2.00%	HTR+
TOMEČEK	Reambulačný protokol	1820	245	lat	novolatinská kurzíva (RKP)	1	Metales Model 4	112996 0	7690	1773	1,54%	4,25%	HTR+
KURHAJCOVÁ	J. M. Hurban	1838/1887	902	slo	kurzíva (RKP)	1	Model J. M. Hurban 50 Model SK+HU	47833 47892	10396 6101	1831 1275	5,10% 5,00%	8,00% 8,70%	Pylaia
KUNEC	Kanonické vizitácie	1754/1756	350	lat	novolatinská kurzíva (RKP	3+	BBBDA – Kanonická vizitácia – 5th	45220	7134	992	0,32%	6,07%	HTR+
BÔBOVÁ	Elenchus librorum	1802/1852	512	lat; ger	novogotická kurzíva (RKP)		Latinský+ Nemecký	45004+ 30919	7596	2316	1,56%	3,08%	HTR+
NIŽNÍKOVÁ MIKUŠKOVÁ	Orbis pictus	1798	94	lat; hun; ger, cze	antikva; kurzíva; fraktúra; švabach (PRT)	tlač	Model 5	44220	4107	1318	0,36%	1,02%	HTR+
KATUŠČÁK[1]	Andrej Kmeť	1860/1908	3000	slo	kurzíva, kurent (RKP)	1	Andrej Kmet Kmet_Py	36009 49016	28672 26749	4703 4395	1,87% 2,10%	5,79% 5,30%	Pylaia
KATUŠČÁK	Lužica	1909	Ca 1000	hsb,dsb	antikva (PRT)	tlač	Lužica_print_1909	42101	10716	1428	0,00%	0,70%	HTR+
KATUŠČÁK	(Varia print)	1960/1910	Ca 500	slo/cze	fraktura_antikva: CZ (PRT)	tlač	Fraktura_antikva:CZ	36550	20805	2252	0,39%	0,44%	HTR+

29 KATUŠČÁK, Dušan, 2022. Umelá inteligencia pomáha sprístupňovať písomné dedičstvo. In: *Knihovna – knihovnická revue*. Praha: Národní knihovna. Roč. 33, č. 2, s. 50 – 77. ISSN 1801-3252. Dostupné: https://knihovnarevue.nkp.cz/archiv/2022-2/recenzovane-prispevky/umela-inteligencia--pomaha-spristupnovat-pisomne-dedicstvo

KAPITOLA 9

VÝKLADOVÝ SLOVNÍK POJMOV A TERMÍNOV

(Automatické rozpoznávanie historických rukopisných a tlačených textov. Upravené a doplnené vysvetlivkami s použitím Transkribus glossary.[1])

Dušan Katuščák

Obrázok 68 Oblak pojmov

Analýza rozloženia. Uplatnenie metódy analýzy obrazu a textovej analýzy, pričom výsledkom tejto analýzy je určenie členenia stránky textu na časti stránky – analýzou sa vyznačujú hlavne bloky textu, horizontálne členenie textu, podstatné, prípadne okrajové, nadbytočné časti obrazu, riadky a základné čiary. Jednotlivé nahraté dokumenty v zbierke majú v nástroji Transkribus Expert Client formu obrázkov, ktoré vznikli v procese snímania (skenovania). Sú to obrázky stránok dokumentov nahratých do platformy Transkribus napríklad vo formáte PDF, JPG, PNG, TIFF. Obrázky je potrebné segmentovať, identifikovať jednotlivé prvky obrázkov. Na účely transkripcie dokumentu je najprv potrebné obrázok rozdeliť na textové oblasti a riadky (TR – Text Regions a Lines). Analýzu rozloženia je možné vykonať niekoľkými kliknutiami a vo väčšine prípadov si úkon nevyžaduje manuálne opravy. To závisí od zložitosti štruktúry vstupného dokumentu. V Transkribus Lite sa analýza rozloženia (segmentácia) spustí automaticky,

1 READ-COOP Transkribus glossary. Dostupné: Transkribus Glossary | READ-COOP (readcoop.eu)

keď sa spustí úloha rozpoznávania textu. Automatická pokročilá analýza rozloženia CITlab vo svojom štandardnom nastavení zvyčajne rozpozná jednu oblasť textu (TR) na obrázku so zodpovedajúcimi základnými čiarami. (Existujú však aj rozloženia, pri ktorých sa odporúča použitie viacerých textových oblastí. Ide o situácie, keď existujú poznámky na okraji alebo poznámky pod čiarou a podobné opakujúce sa prvky. Pokiaľ sú tieto textové oblasti, ktoré sa líšia obsahom a štruktúrou, obsiahnuté v jednej textovej oblasti, analýza rozloženia jednoducho počíta riadky zhora nadol. Toto poradie čítania nezohľadňuje, kam text skutočne patrí z hľadiska obsahu, ale len to, kde sa na stránke graficky nachádza. Oprava automaticky vygenerovaného, ale neuspokojivého poradia čítania môže byť časovo náročná. Problému možno ľahko predísť vytvorením niekoľkých textových oblastí (TR).

Archívne fondy a zbierky. Historické rukopisné**,** prípadne strojopisné dokumenty na transkripciu sa nachádzajú prevažne v archívoch. Historické tlačené dokumenty sa nachádzajú hlavne v knižniciach, ale aj u iných právnických alebo fyzických osôb. Na usporiadanie archívnych fondov sa u nás používa *Klasifikačná schéma archívnych fondov a zbierok štátnych archívov na Slovensku*. Na najvyššej úrovni majú archívy spravidla svoje *zoznamy archívnych fondov a zbierok*. Tieto zoznamy obsahujú všeobecné atribúty fondu a zbierky: názov archívneho fondu/zbierky, časové rozpätie, rozsah veľkosti archívneho fondu/zbierky v bežných metroch, prístupnosť a typ archívnej pomôcky. Výber konkrétnych dokumentov na transkripciu a výskum záleží na erudícii výskumníka, pretože rozsah a hĺbka spracovania fondov a zbierok je rôzna.

Canvas (menu úprav v Transkribus Expert Client). Spustenie automatickej analýzy (segmentácie) rozloženia stránky a textu neposkytuje vždy vyhovujúce výsledky. Niekedy sú preto potrebné manuálne korekcie rozloženia. V ponuke Canvas, ktorá sa nachádza spravidla na ľavej strane stránky dokumentu, sa nachádzajú potrebné voľby, ako je ohraničenie bloku textu (TR – Text Regions), pridanie riadku (L – Lines), pridať základnú linku (BL – Base Lines), pridať slovo (W – Word), pridávanie rôznych častí (tabuľky, reklamy, schémy, grafy, grafiky atd.). V ponuke Canvas je môžné tiež zmeniť existujúce tvary.

CER (Character Error Rate). 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 na získanie výsledku *Ground Truth*. Ide teda o chyby v porovnaní s presným, referenčným textom. Vzorec na výpočet CER je: 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. Považuje za potvrdené a overené konštatovanie, že: a) ak je hodnota chybovosti *znakov* CER 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*.

Cvičenie modelu. Pomocou nástroja Transkribus Expert Client je môžné cvičiť (trénovať) model rozpoznávania rukopisného textu, aby bolo možné transkribovať automaticky zbierky dokumentov. Model je výsledkom cvičenia, preto je pri jeho tvorbe potrebné cvičiť tak, aby stroj rozpoznal určitý štýl písania v zobrazovaných obrázkoch dokumentov a poskytol ich viac-menej presný prepis. Na cvičenie (TRAIN) modelu je potrebných 5 000 až 15 000 slov (približne 25 – 75 strán) prepísaného materiálu. Prepis sa získa manuálnym prepisom riadok po riadku presne podľa predlohy. Prepis si možno uľahčiť použitím už prepísaných a dostupných dokumentov alebo postupovať pri príprave cvičného súboru s použitím základného súboru. Pri práci s tlačeným textom sa zvyčajne vyžaduje menšie množstvo cvičných údajov ako pri rukopisoch. Použitím základného modelu je možné znížiť množstvo požadovaných cvičných dát. Ako základný model sa môže použiť buď jeden z verejne dostupných modelov PyLaia v Transkribe, ktorý by mohol byť vhodný pre naše dokumenty, alebo jeden z našich vlastných modelov, ktoré sme už predtým cvičili.

DocScan. Open source aplikácia pre Android navrhnutá pre ScanTent. Identifikuje strany dokumentu v živom náhľade a robí snímky v dostatočnej kvalite na transkripciu. V automatickom režime nasníma obrázok po otočení stránky. Umožňuje rýchlo snímať knihy alebo dokumenty bez interakcie s mobilom. Obrazovku smartfónu je možné zdieľať na obrazovke počítača a vzdialene ovládať smartfón napríklad cez TeamViever. Vďaka spoločnosti ifunplay a aplikácie DocScan možno teraz ScanTent používať aj s operačným systémom iOS v iPhonoch. Držiak na vrchu ScanTent umožňuje umiestnenie smartfónu a optimálny pozorovací uhol a konštantnú vzdialenosť. Ak denné svetlo nestačí, biele LED pásiky poskytujú rovnomerné osvetlenie, ktoré maximalizuje kvalitu obrazu. Poznámka: Francúzska národná knižnica používa 40 zariadení ScanTent.

Dokument (Document). V štruktúre systému Transkribus Expert Client je dokument zaradený do nejakej zbierky. Dokument môže byť presunutý do inej existujúcej zbierky. Základné metadáta k dokumentu sú: jedinečný číselný identifikátor, názov dokumentu, meno osoby, ktorá nahrala dokument do zbierky v Transkribe, dátum a čas nahratia do zbierky, meno zbierky, do ktorej dokument patrí. Dokument je možné zobraziť vo forme „Overview“ s jednotlivými stranami a grafickým rozlíšením stavu stránky (napr. Ground Truth, In progress, Done, Final). Vo forme „Layout“ sú viditeľné texty transkripcie strán, riadky textu, poradie čítania riadkov strojom, identifikátor riadka a koordináty umiestnenia elementov v riadku.

Export. Ak chceme pracovať so svojimi obrázkami a prepismi mimo Transkribu, môžeme svoje dokumenty exportovať do bežnejších formátov, ako sú DOCX, PDF, EXCEL, PageXML, TEI-XML alebo TXT. Možnosti zahŕňajú export celých strán, obrázkov, textu alebo štrukturálnych prvkov. Exportovať je možné do nejakého adresára na lokálnom počítači alebo exportovať na server Transkribus, z ktorého nám príde oznámenie po skončení exportu.

Formát JPG, JPEG. Najrozšírenejší je formát, ktorý sa vyskytuje s príponou .jpg, .jpeg alebo .JPG, .JPEG. Medzi nimi nie je žiadny rozdiel. V tomto formáte ukladajú

súbory všetky fotoaparáty aj mobilné zariadenia, ak používame napríklad DocScan. V niektorých aparátoch je možné voliť jeden formát alebo snímanie v dvoch formátoch JPG a RAW, ARW. Výhodou formátu JPG je, že sa obrázok dá zobraziť prakticky v každom zariadení – v mobilnom telefóne, televízore alebo vo webovom prehliadači. Zaberá málo miesta na disku, je úsporný, pretože ide o kompresiu so stratou. Nevýhodou tohto formátu je, že každou úpravou obrázok stráca kvalitu pri každom uložení. V projektoch transkripcie používame na snímanie mobilnými zariadeniami formát JPG na archivovanie a v transkripcii spravidla pracujeme s derivovaným formátom PDF.

Formát RAW. Znamená, že nasnímaný súbor je „surový", nespracovaný a dáta nie sú komprimované. Dáta v tomto formáte sú veľmi veľké a na ich spracovanie je potrebný špeciálny softvér, napríklad komerčný Zoner Photo Studio alebo open source FastStone Image Viewer. Výsledné obrázky majú vysokú kvalitu a sú po úprave hodné na kvalitné editovanie.

Formát TIFF. Vyskytuje sa s príponami .tiff, tif. Pri ukladaní do tohto formátu spravidla nedochádza ku kompresii dát. Ak áno, tak ide o bezstratovú kompresiu aj pri opakovanom ukladaní. Súbor zachováva maximum informácií z formátu RAW pri editácii. Nevýhodou je veľkosť súborov vo formátoch TIFF. V profesionálnych projektoch digitalizácie je formát TIFF najvhodnejší na dlhodobé archivovanie.

Formáty obrázkov. Snímky je možné tvoriť, ukladať a upravovať v rôznych formátoch. Najčastejšie ide o súbory vo formátoch RAW a JPG. Z hľadiska úprav fotografií je dôležitý formát TIFF.

Fuzzy vyhľadávanie (Fuzzy Search). Technika vyhľadávania, ktorá umožňuje nájsť okrem presných zhôd s hľadaným výrazom aj podobné slová. Pomocou expertného klienta môžete určiť fuzzy vyhľadávanie, t. j. že nie všetky znaky hľadaného slova sa musia zhodovať. To sa môže hodiť, keď sú prijateľné alternatívne pravopisy.

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 krajinách hovoriacich po nemecky.

HTR+ a PyLaia. Softvér HTR+ spoločnosti Transkribus zatiaľ nemôže okamžite spustiť spoľahlivý automatický prepis, ale najprv musí byť vyškolený na konkrétny typ písma a rukopisu. HTR+ bol do konca roku 2022 aj vo výskume Skriptor používaný ako hlavný stroj na rozpoznávanie rukopisného textu vyvinutý tímom CITlab na Univerzite v Rostocku. Transkripčný mechanizmus je založený na TensorFlow. Namiesto HTR+ je v súčasnosti v Transkribe dostupný nástroj PyLaia.

Import dokumentov (Upload). Po vytvorení zbierky je potrebné v Transkribe nahrať

dokumenty (Upload). Potom je možné spustiť nástroje, ako je analýza rozloženia (segmentácia) alebo rozpoznávanie textu (transkripcia). Údaje v Transkribe sú vždy súkromné a prístupné iba jednotlivým používateľom. Vlastník zbierky (owner) môže umožniť prácu aj iným používateľom (users) s oprávneniami, ktoré im pridelí (owner, editor, transcriber, reader).

ISAD(G) (General International Standard Archival Description). Medzinárodný štandard, ktorý definuje zoznam prvkov a pravidiel na popis archívov a popisuje druhy informácií, ktoré musia a mali by byť zahrnuté v takýchto opisoch. Vytvára hierarchiu popisu, ktorá určuje, aké informácie by mali byť zahrnuté na akej úrovni. V súvislosti s výskumom a experimentmi s transkripciou archívnych dokumentov považujeme za vhodné, aby boli transkribované fondy, zbierky a dokumenty popísané na štandardnej úrovni. Tento štandard poskytuje rámec pre spoločný prístup a nie rigidný formát.

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 0 a 1 (0 = najnižšia a 1 = najvyššia), aby sa zhodnotila kvalita výsledkov hľadania.

Model. V platforme Transkribus je model entita, ktorá je výsledkom použitia softvéru strojového učenia a umelej inteligencie a hlbokých neurónových sietí na rozpoznávanie historických rukopisných a tlačených textov. Platforma Transkribus umožňuje používateľom cvičiť model rozpoznávania textu rukou (HTR+, PyLaia) na automatické spracovanie zbierky dokumentov. Model je potrebné cvičiť tak, aby rozpoznal určitý štýl písania zobrazovaním obrázkov dokumentov a umožnil ich presný prepis. Podľa typu textu môžu používatelia na transkripciu použiť verejne dostupný model alebo vytvoriť vlastný model, prípadne cvičiť vlastný model s použitím základného modelu. Na cvičenie modelu použijeme voľbu nástroje (Tools). V časti cvičenie modelov (Model training) zvolíme cvičiť nový model (Train a new model). Zobrazí sa okno Model Training. V predvolenom nastavení je vybratý motor „PyLaia HTR".

Oblasť textu; Blok textu (Text region TR). Ak chceme vygenerovať prepis HTR, musíme dokumenty rozdeliť na textové oblasti, riadky a základné čiary. V predvolenom nastavení je oblasť textu obdĺžnik, ktorý obklopuje všetok ručne písaný text obsiahnutý v obrázku. Je však možné upraviť textovú oblasť podľa všeobecného rozloženia pridaním kontrolných bodov, čím sa vytvorí polygón.

OCR (Optické rozlišovanie písma; Optical Character Recognition) Optické rozpoznávanie znakov alebo optická čítačka znakov (OCR) je elektronická alebo mechanická konverzia obrázkov ručne písaného alebo vytlačeného textu na strojovo kódovaný text, či už z naskenovaného dokumentu alebo fotografie.

Polygóny (Polygons). Historické dokumenty majú niekedy zložité usporiadanie a pozostávajú z rôznych rozložení, čo môže viesť k problémom s poradím čítania prvkov textu. Pri komplikovaných rozloženiach si rýchlo všimneme, že ručne nakreslené textové oblasti sa môžu prekrývať. Tento problém sa dá ľahko vyriešiť úpravou pravouhlých oblastí textu, pridaním bodov a tým vytvorením polygónov.

Poradie čítania. V systéme Transkribus Expert Client poradie čítania zobrazuje na segmentovanej stránke to poradie, v ktorom bude stroj transkripcie čítať riadky textu na obrázku stránky. Toto poradie čítania sa vytvára automaticky počas analýzy rozloženia, ale možno ho neskôr zmeniť aj manuálne. Pri automatickej analýze rozloženia je poradie čítania určené súradnicami riadkov na obrázku: horný riadok, ktorý je najviac vľavo, je číslo jedna atď. Dôležité je vedieť, že poradie čítania nie je relevantné pre samotné školenie, ale môže sťažovať čítanie transkribovanej strany. Ak sa má prepis exportovať a ďalej použiť na vydanie, tak poradie čítania je potrebné zadať správnym spôsobom, aby bol text v správnom poradí. Dá sa to jednoducho urobiť zapnutím poradia čítania na karte *viditeľnosť tvaru*. Všetky riadky tak zobrazujú kruh s číslom, ktoré označuje ich polohu na stránke dokumentu. Kliknutím na tieto krúžky sa otvorí okno s textovým editorom, kde je možné priradiť nové, správne čísla. Táto funkcia je užitočná najmä v dokumentoch s náročným rozložením, kde sa poradie riadkov neriadi bežnými pravidlami.

Presnosť výpočtu. Presnosť modelu je možné merať na konkrétnych stránkach z cvičných a overovacích sád pomocou funkcie *presnosť výpočtu* (Compute accuracy...) na karte *nástroje* (Tools). Na tento účel je najprv potrebné generovať transkripciu HTR. Na porovnanie textových verzií sú potrebné dva transkribované súbory: *referencia* (Reference) – správny text a *hypotéza* (HTR, transkribovaný text). Ako *referencia* sa vyberie verzia stránky, ktorá bola správne prepísaná, teda *„základná pravda“* (Ground Truth), čo je manuálny prepis čo najbližšie k pôvodnému textu. Na získanie najvýznamnejšej hodnoty by bolo najlepšie použiť stránky zo vzorového súboru, ktoré neboli použité v tréningu, a preto sú pre model nové. Použitie stránok z overovacieho súboru je tiež možnosťou, aj keď nie ideálnou. Použitie stránok z cvičného súboru nie je vhodné, pretože to prinesie nižšie hodnoty CER , ako v skutočnosti sú. Ako *hypotézu* vyberieme verziu, ktorá bola automaticky vygenerovaná pomocou modelu HTR, na ktorej chceme vidieť, aký dobrý je výsledok.

Princípy popisu ISAD(G) sa riadia štyrmi všeobecnými zásadami: *1) Opis od všeobecného po konkrétny* – viacúrovňový opis sa začína od všeobecnej úrovne opisu, ktorá je zvyčajne fondmi, a pokračuje do podrobnejších úrovní, ako sú podfondy, séria, súbor, položka atď. Táto hierarchická štruktúra musí byť reprezentovaná a správne definovaná v archívnom opise. *2) Informácie relevantné pre úroveň opisu* – informácie na každej úrovni opisu sa musia týkať len archívnej jednotky opísanej na tejto úrovni. *3) Prepojenie popisov* – každá archívna jednotka musí byť prepojená so svojou nadradenou úrovňou v rámci hierarchie a jej úroveň musí byť explicitná. *4) Neopakovanie informácií* – aby sa zabránilo opakovaniu, všeobecné informácie spoločné pre skupinu sa musia deklarovať na najvyššej možnej úrovni. Podúrovne musia zase obsahovať spoločné informácie, ktoré sa vzťahujú na jej nižšie úrovne.

Prvky popisu ISAD(G). Štandard definuje dvadsaťšesť dátových údajov popisu archívnych fondov, zbierok a dokumentov. Vo výskume Skriptor sa odporúča použiť na popis zbierok a dokumentov tieto pravidlá s návestiami v uvedenom poradí. Popisy je ďalej možné použiť pri editovaní zbierok, špecifikácii metadát v Transkribus Expert Client a na prezentáciu transkribovaných zbierok v nástroji read&search. V prvej oblasti ***1) Vyhlásenie o totožnosti*** týchto 6 povinných údajov identifikujúcich fond, zbierku, dokument: *2) Referenčné kódy*: Prvky používané na jednoznačnú identifikáciu jednotky opisu: kód krajiny, kód úložiska, špecifický miestny referenčný kód/kontrolné číslo/iný jedinečný identifikátor. *3) Titul*: Názov jednotky opisu. *4. Dátum*: Dátumy vytvorenia záznamu počas vedenia záležitostí alebo dátumy vytvorenia dokumentu. *5) Úroveň popisu*: Úroveň jednotky opisu v rámci hierarchie. *6. Rozsah a médium* jednotky opisu: Fyzikálny alebo logický rozsah a médium jednotky opisu. Ďalej sú oblasti popisu. V oblasti ***2. Kontext*** je povinný len údaj *Meno*: Tvorca jednotky popisu. Ďalšie údaje v oblasti kontext sú: *Administratívne/biografické dejiny*: Biografické alebo administratívne podrobnosti týkajúce sa tvorcov jednotky opisu; *Archívna história*: Príslušné historické informácie o jednotke opisu; *Bezprostredný zdroj akvizície alebo prevodu*: Zdroj získania materiálu. V oblasti ***3. Obsah a štruktúra*** sú údaje: **Rozsah pôsobnosti a obsah:** Zhrnutie rozsahu a obsahu relevantného pre úroveň opisu; **Informácie o hodnotení, zničení a plánovaní:** Zhodnotenie, zničenie a plánovanie činností vykonaných alebo plánovaných pre jednotku opisu. **Časové rozlíšenie:** Plánované dodatky k jednotke opisu. V oblasti ***4. Podmienky prístupu a používania*** sú údaje**: Podmienky upravujúce prístup:** Informácie o právnom postavení, ktoré môžu ovplyvniť prístup k jednotke opisu; **Podmienky, ktorými sa riadi reprodukcia:** Podmienky na reprodukciu jednotky opisu po vytvorení; **Jazyk/skripty materiálu: Jazyky, skriptá** a systémy symbolov používané v jednotke opisu; **Fyzikálne vlastnosti a technické požiadavky:** Príslušné fyzické podmienky, softvérové a hardvérové požiadavky na prístup k jednotke opisu a jej uchovávanie; **Hľadanie pomôcok:** Nájdenie pomôcok použiteľných pre jednotku opisu. V oblasti ***5. Spojené materiály*** sú údaje**: Existencia a umiestnenie originálov:** Informácie o existencii alebo zničení pôvodnej jednotky opisu; **Existencia a umiestnenie kópií:** Informácie o existencii a dostupnosti kópií jednotky opisu; **Súvisiace jednotky opisu:** Informácie o jednotkách opisu súvisiacich s pôvodom alebo inými asociáciami s jednotkou opisu; **Poznámka k uverejneniu:** Publikácie, ktoré sa týkajú alebo sú založené na použití, štúdiu alebo analýze jednotky opisu; V oblasti ***6. Poznámky*** *sú* informácie, ktoré sa nezmestia do žiadnej z predchádzajúcich oblastí; V oblasti ***7. Kontrola*** sú údaje**: Poznámka archivára:** Informácie o tom, kto a ako pripravil popis. **Pravidlá alebo dohovory:** Protokoly, na ktorých je opis založený. **Dátum(-y) popisu:** Dátumy vytvorenia a revízie.

PyLaia. Od roku 2022 preferovaný 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 Transkribe.

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. Výskum bol predtým financovaný ako súčasť projektu tranScriptorium. Tento projekt získal finančné prostriedky zo 7. rámcového programu Európskej únie pre výskum, technologický rozvoj podľa dohody o grante č. 600707.

Read&search. Platforma Transkribu, ktorý sprístupňuje online dokumenty zo zbierky vytvorenej v platforme Transkribus Expert Client. Webové rozhranie bohaté na funkcie je ideálne na sprístupnenie historických dokumentov a vyhľadávanie na webe.

READ-COOP. Združenie na udržateľnosť a vývoj platformy Transkribus. 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. V READ-COOP sa kupujú kredity. Nejde o zisk združenia, ale o príjem, ktorý sa používa na výskum, vývoj a infraštruktúru. V projekte SKRIPTOR „spotrebujeme“ asi 10 000 kreditov (1944 €).

Riadkové oblasti (Line Regions LR). Oblasti, ktoré sa nachádzajú v rámci bloku textu a možno ich opísať ako mnohouholníky, v ktorých je všetok ručne písaný text v riadku. Keďže nemajú pre proces transkripcie bezprostredný význam, riadkové oblasti by sa nemali opravovať. Ak sa niečo má zmeniť v rozložení riadkov dokumentu, vždy to treba urobiť na základnej úrovni (baseline). Základná čiara (baseline) by mala prebiehať pozdĺž spodnej časti textového riadku, písmená by na nej mali sedieť a zostupne smerovať nižšie. Čiarové oblasti sa prispôsobia automaticky, keď niečo zmeníte na základnej úrovni. Zobrazí sa vyskakovacie okno s otázkou, či by ste chceli zmeniť aj nadradený riadok, čo treba potvrdiť.

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.). V roku 2017 sme pracovali s verziou Transkribus Expert Client v1.3.7. V októbri roku 2022 bola k dispozícii verzia 1.22.0.

Snímanie. Jeden z procesov digitalizácie. Vykonáva sa pomocou vhodného technického zariadenia na digitalizáciu, akými sú zariadenia na zachytenie digitálneho obrazu (digitálne fotoaparáty a kamery, skenery na knihy, dokumenty alebo mikrofilmy, audio- a videohardvér) pripojené na vhodnú počítačovú platformu. Je

možné rozlíšiť dve rôzne metódy snímania: *skenovanie* a *fotografovanie*, používanie *digitálnych kamier/fotoaparátov*, *mobilných telefónov*. V súlade s predpisom Ministerstva vnútra SR archív *„d) umožňuje snímanie archívnych dokumentov a priestorov archívu klasickou kamerou a digitálnou kamerou (ďalej len „kamera“) a fotografickou technikou“*. Na účely automatickej transkripcie, pokiaľ je to možné, použijeme dokumenty nasnímané profesionálnymi skenermi a obrazmi v najvyššej dosiahnuteľnej kvalite. Minimálna kvalita skenovania by mala byť 300 DPI. Nakoľko pri historických rukopisoch ide *de facto* o grafiku, je vhodné skenovať vo vyššej kvalite. Pre platforme Transkribus je možné snímať dokumenty do fomátu veľkosti A3 zariadením ScanTent so softvérom DocScan.

Spracovanie archívnych zbierok. Archívne fondy a zbierky často predstavujú obrovské množstvá dokumentov. Na Slovensku štátne archívy uchovávajú asi 200 kilometrov dokumentov. Podľa výskumu Medzinárodnej archívnej asociácie (ICA) z roku 2020 nemá 3 % archívov zbierky vôbec spracované a 50 % archívov má zbierky spracované na základnej úrovni. Výskum tiež ukázal nízku mieru využívania medzinárodných štandardov odporúčaných ICA. Ide o štandardy: ICA ISAD(G), ISDF, ISAAR (CPF) a ISDIAH. Najrozšírenejší je štandard ISAD(G).

Statusy transkripcie. Rôzne stavy spracovania strany: *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 strany sa môžu priradiť rôzne stavy v závislosti od toho, aký pokrok sa na nich už dosiahol.

Tabuľky. Tlačené a ručne kreslené tabuľky sú bežné v historických dokumentoch všetkých typov. V súčasnosti sa tabuľky musia v Transkribe kresliť ručne pomocou editora tabuliek. Technológia, ktorá umožní automatické rozpoznávanie tabuliek, je vo vývoji. V súčasnosti ide v práci s tabuľkami o poloautomatický proces. Na účely transkripcie je najprv potrebné manuálne vytvorenie štruktúry tabuľky v Transkribe a prepis textu, ktorý tabuľka obsahuje. Ak majú tabuľky v dokumente rovnakú štruktúru na viacerých stranách, je možné schému pripravenej štruktúry tabuľky použiť na dávkové rozpoznanie ďalších strán s tabuľkami. Ak teda majú viaceré strany rovnakú štruktúru tabuľky alebo šablónu tabuľky, pripraví sa manuálne tabuľka len pri prvom výskyte tabuľky a potom sa distribuuje na ďalšie strany pomocou súpravy nástrojov *nomacs*. Na transkripciu tabuliek sa najprv vytvoria textové oblasti (TR) pre všetky informácie, ktoré nepatria do tabuľky. Týka sa to informácií v hornej, spodnej časti alebo po stranách stránky, ktoré zjavne nie sú súčasťou tabuľky, ako napríklad: čísla strán, čísla riadkov, termíny, akékoľvek iné označenia alebo anotácie. Následne sa vytvoria textové oblasti (TR) pre jednotlivé bunky tabuľky, horizontálne a vertikálne čiary a koriguje sa text v bunkách tabuľky na strane. Grafickú schému tabuľky, ohraničenie tabuľky a buniek je možné použiť na ďalšie rovnaké strany s tabuľkami. Bunky sa ohraničujú pomocou volieb v nástroji „Cell borders“.

Tagovanie štruktúry. V štruktúre systému Transkribus Expert Client je možné pomocou funkcie štrukturálneho značkovania vo funkcionalite *metadáta* označiť, „značkovať" (mark-up) prvky štruktúry dokumentov. Navyše je možné cvičiť modely tak, aby automaticky rozpoznali štruktúru dokumentov. Pridaním tagov, teda štrukturálnych značiek sa vytvoria cvičné dáta pre tento proces. Nie je potrebné označovať každý prvok dokumentov – stačí sa zamerať na označenie sekcií, ktoré nás zaujímajú. Rozhranie štrukturálneho označovania v Transkribe umožňuje rozdeliť dokumenty do štruktúrnych sekcií, ako sú odseky, nadpisy alebo čísla strán, pridať prispôsobené kategórie značiek pre vaše individuálne potreby a v budúcnosti použiť tieto štrukturálne informácie na cvičenie modelu.

TensorFlow a PyTorch. Stroje HTR+ a PyLaia vychádzajú zo softvéru TensorFlow a PyTorch, čo sú bezplatné softvérové knižnice s otvoreným zdrojovým kódom pre strojové učenie a umelú inteligenciu . TensorFlow slúži na rozpoznávanie obrázkov. Poskytuje štandardný postup, ktorý zahŕňa triedenie pixelov obrázka, získanie vlastností pixelov, trénovanie obrázka, trénovanie modelu a testovanie modelu oproti vstupom. TensorFlow je možné použiť aj na detekciu jazyka, preklad, rozpoznávanie vzorov rukopisu atď. Ich najbežnejšie uplatnenie je v bankách a poisťovniach pri odhaľovaní podvodov. Dá sa použiť v celom rade úloh, ale špeciálne sa zameriava na cvičenie a odvodzovanie hlbokých neurónových sietí. TensorFlow bol vyvinutý tímom Google Brain na interné použitie Googlom vo výskume a výrobe. Pôvodná verzia bola vydaná pod licenciou Apache 2.0 v roku 2015. Google vydal aktualizovanú verziu TensorFlow s názvom TensorFlow 2.0 v septembri 2019. TensorFlow je možné použiť v širokej škále programovacích jazykov vrátane Pythonu, JavaScriptu, C++ a Java. Táto flexibilita sa hodí pre celý rad aplikácií v mnohých rôznych sektoroch. Jednou z aplikácií TensorFlow a PyTorch je Transkribus so strojmi HTR+ a PyLaia. PyTorch umožňuje vyspelejšie rozpoznávanie textu. Na trénovanie modelu rozpoznávania textu založeného na umelej inteligencii sa používa rekurentná neurónová sieť (RNN) a PyTorch. Ďalšie podobné aplikácie zahŕňajú detekciu rukopisu, rozpoznávanie vzorov atď.

Transkribus Expert Client. Samostatná profesionálna verzia Transkribu s plným výkonom platformy Transkribus.

Transkribus Lite. Verzia prehliadača Transkribus. Automaticky transkribuje a umožňuje pohodlnú úpravu historických dokumentov. V Transkribus Lite je možné cvičiť vlastné modely AI v nejakom prehliadači. V prehliadačoch osobných počítačov a smartfónov je možné prezerať a upravovať zbierky z *Transkribus Expert Client*. Mnohé z funkcií klienta *Transkribus Expert Client* môžu byť použité aj v *Transkribus Lite*. V Transkribus Lite sa analýza rozloženia (segmentácia) spustí automaticky, keď sa spustí úloha rozpoznávania textu.

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. Platforma integruje nástroje vyvinuté výskumnými skupinami v celej Európe vrátane skupiny na 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).

Transkripcia (prepis). Podľa wikipédie: a) v užšom zmysle: písomné vyjadrenie (vyslovovaných alebo cudzím grafickým systémom napísaných) slov a textov z hľadiska ich výslovnosti prostriedkami určitého grafického systému (v najužšom zmysle len takéto písomné vyjadrenie slov a textov napísaných cudzím grafickým systémom); b) v širšom zmysle: bod a) plus transliterácia; c) v najširšom zmysle: vyjadrenie výrazu jedného jazyka v inom grafickom systéme ako v grafickom systéme, v ktorom sa tento jazyk obyčajne zapisuje, resp. v ktorom je už zapísaný (táto definícia zahŕňa okrem bodu b) napr. prevod slovenského textu do Braillovho písma, prevod posunkovej reči do písanej notácie a [pravdepodobne aj] moderný prepis textu archiválie). V platforme Transkribus sa používa termín transkripcia vo význame prepisu rukopisného alebo tlačeného historického textu v určitom jazyku a automatický prepis textu v tom istom jazyku. Napríklad rukopis v maďarčine sa prepisuje pomocou znakovej sady tlačenej latinky. Nejde teda o prepis medzi jazykmi, ale o prepis v rámci jedného jazyka.

Transliterácia. Odborný alebo vedecký prepis; odborná alebo vedecká transkripcia, zriedkavo prepísmenkovanie sa v jazykovede definuje: a) po písmenách uskutočňované „pretlmočenie" textov či slov zapísaných jedným grafickým systémom prostredníctvom iného grafického systému; b) grafický prepis cudzojazyčného textu nahradzujúci písmená jednej abecedy písmenami druhej abecedy bez prihliadnutia k fonetickej hodnote, takže je možný spätný prepis; c) prevod z jedného grafického systému do druhého, pričom každému písmenu jedného grafického systému zodpovedá vždy písmeno druhého systému (rovnaké písmeno alebo spojenie písmen), takže je možný aj jednoduchý spätný prevod do jazyka originálu. Podobná definícia: Prepis z jedného písma do druhého, pri ktorom sa zachováva jednoznačnosť zápisu medzi písmenami obidvoch abecied. Transliteráciou sa podľa Pravidiel slovenského pravopisu (2013) rozumie napríklad prepis zo slovanských jazykov používajúcich cyrilské písmo do slovenčiny (do latinského písma používaného v slovenčine). Teda prepis znakov, písiem z jedného jazyka (napríklad ruštiny) do iného jazyka (slovenčiny) alebo prepis z iných grafických sústav do latinky. Pravidlá transliterácie cyrilských písmen zo súčasných slovanských jazykov, ktoré používajú cyriliku, určujú slovenské technické normy. Prepisy z najdôležitejších jazykov Ďalekého východu, a to z čínštiny, japončiny a kórejčiny upravujú prepisy, v ktorých sa čo najvernejšie zachytáva ich zvuková podoba. V platforme Transkribus sa viac konvenčne ako presne používa termín transkripcia. Termíny transliterácia a transkripcia sa používajú často v rovnakom význame.

Verejné modely transkripcie. Modely Transkribu, ktoré je možné použiť na podobné dokumenty. Pre každý model je uvedený krátky opis cvičného materiálu, pre ktoré jazyky môže byť model užitočný a kto ho vytvoril a cvičil. Cieľom je sprístupniť používateľom Transkribu čoraz viac modelov, aby mohli ťažiť z kooperácie a sieťového efektu a šetriť

prácu a čas. V novembri 2022 bolo dostupných 97 verejných modelov, napríklad: nemecký kurent, noviny, časopisy, rôzne tlače a rukopisy; viacjazyčný model pre tlače v rôznych jazykoch (holandčina, angličtina, fínčina, francúzština, nemčina, švédčina); všeobecný model pre francúzske rukopisy, nemecká bastarda 15. st.; dánska fraktúra a historické rukopisy a strojopisy; holandské rukopisy a tlače; estónske rukopisy; fínske noviny a rukopisy; francúzske rukopisy a tlače; hlaholika; latinčina; neolatinčina; ruština; španielske rukopisy a tlače a pod.

Verzie. Pri práci so systémom Transkribus Expert Client sa pri každom spustení úlohy alebo uložení dokumentu vytvorí nová verzia dokumentu. Výhodou je, že sa vždy môžete vrátiť k starším verziám a pokračovať v práci na nich, čo sťažuje stratu údajov v Transkribe. Navyše je možné porovnávať verzie s nástrojom Compute Accuracy v Transkribe. Pri verziách jednotlivých stránok je vždy informácia o stave (statuse) strany, používateľovi, dátume zmeny, nástroji zmeny a identifikátoroch.

Virtuálna klávesnica. Editačný nástroj Transkribus Expert Client, ktorý umožňuje pridávať špeciálne znaky a Unicode (ISO 10646), ktoré nie sú dostupné na bežnej klávesnici. Nachádza sa v poli textového editora v spodnej časti okna expertného klienta. Pomocou tlačidla *upraviť…* je možné pridávať skratky pre často používané znaky a pridávať nové znaky Unicode. Ak je potrebné vytvoriť skratku, stačí ju zadať do stĺpca *skratka* a na pridanie nových znakov Unicode použiť zelené tlačidlo plus.

Vyhľadávanie. V dokumentoch, ktoré boli v Transkribe transkribované pomocou HTR-modelu, je možné vyhľadávať pomocou kľúčových slov pomocou fulltextového vyhľadávania (Solr). Systém umožňuje (pravdepodobnostné) „fuzzy vyhľadávanie“ (Fuzzy Search), čo je vyhľadávacia technika, ktorá umožňuje nájsť podobné slová okrem presných zhôd pre hľadaný výraz. Pomocou expertného klienta môžete zadať fuzzy vyhľadávanie, t. j. že nie všetky znaky hľadaného slova sa musia zhodovať. Text by mal byť okamžite k dispozícii na vyhľadávanie. Indexuje sa vždy iba posledná verzia každého prepisu. Vyhľadávanie je možné po zadaní jedného slova, viacerých slov alebo presnej vety. Vyhľadávanie je možné podľa tagov (značiek) v dokumentoch v zbierke, dokumente, na strane, riadku, s voľbou veľkých a malých písem. Originálne vyhľadávanie umožňuje metóda KWS (The Keyword Spotting). V zbierke je možné hľadať podľa ID zbierky, ID dokumentu, názvu zbierky, popisu a autora zbierky.

WER (Word Error Rate). Hodnota chybovosti slov v transkripcii.

Základná čiara (Baseline). Najdôležitejší referenčný bod na rozpoznávanie textu. Popisuje polyčiaru, ktorá sa tiahne pozdĺž spodnej časti rukou písaného textového riadku. Segmentáciu textu na riadky a základné čiary je možné vykonať automaticky pomocou CITlab Advanced LA. Pri zložitých rozloženiach a v závislosti na konkrétnom písme v rukopisoch sa však môžu vyskytnúť prípady, keď je potrebné vykonať niektoré manuálne opravy. Základná čiara by mala prebiehať pozdĺž spodnej časti textového riadku, písmená by na nej mali sedieť a zostupne smerovať nižšie. Základná čiara pozostáva z jednotlivých bodov, ktoré je možné nastaviť pri manuálnej úprave

sekvencie; nastavenie sa dokončí dvojitým kliknutím alebo voľbou Enter v poslednom bode. Základné línie je možné nakresliť aj vertikálne. Na obrázku a dokonca aj v textovej oblasti je možné tiež kombinovať rôzne smery čiar (napr. typické pohľadnicové rozloženie). Ak sa vykonávajú zmeny na linkách, je dôležité, aby sa vždy robili na základných čiarach, pretože pre každý riadok v dokumente je na pozadí aj oblasť čiary. Dajú sa zobraziť pomocou tlačidla *viditeľnosti položky*. Tieto riadkové regióny sa nesmú meniť, automaticky sa prispôsobia, keď sa niečo zmení na základnej úrovni. Zobrazí sa vyskakovacie okno s otázkou, či by ste chceli zmeniť aj nadradený riadok, čo treba potvrdiť.

Základný model (Base model). Ak tvoríme vlastné, generické modely HTR, tak nepracujeme so základnými modelmi. Pri cvičení so základnými modelmi je však každé cvičenie pre model založené na existujúcom modeli, t. j. na *základnom modeli*. Toto je spravidla posledný model HTR, ktorý bol vyškolený v nejakom projekte. Základné modely si „pamätajú“ to, čo sa už „naučili“. Preto každé nové školenie „teoreticky“ zlepšuje kvalitu novotvoreného modelu. Nový model sa učí od svojho predchodcu a stáva sa tak lepším a lepším. Preto je školenie so základnými modelmi obzvlášť vhodné aj pre veľké generické modely, ktoré sa neustále vyvíjajú počas dlhého časového obdobia. Ak chceme vykonať školenie so základným modelom, jednoducho si v cvičnom nástroji vyberieme konkrétny základný model – okrem obvyklých nastavení. Potom na karte údaje modelu HTR vložíme cvičný súbor a overovací súbor základného modelu, ako aj nový cvičný a overovací súbor. Okrem toho môžeme pridať ďalšie nové strany Ground Truth a začať s cvičením.

Zálohovanie a archivovanie. V procesoch snímania je nevyhnutné zvoliť metódu zálohovania a archivovania zdrojových obrázkov a ich derivátov. Základné pravidlo o zálohovaní vyžaduje urobiť najmenej tri kópie na dva rôzne nosiče a jednu – archívnu zálohu mať na vzdialenom mieste. Každá snímka by mala mať aspoň dve kópie, a to na dvoch rôznych úložiskách, napríklad na SD karte, disku, externom disku, digitálnom repozitári.

Zbierka (Collection). V štruktúre systému *Transkribus Expert Client* sú dva kľúčové prvky: *zbierky* a *dokumenty*. Zbierka je nadradená dokumentu. Dokumenty sú usporiadané do tzv. zbierok. Zbierky možno chápať ako priečinky obsahujúce dokumenty. Zbierky sa zvyčajne tvoria podľa nejakého konkrétneho projektu. Napríklad všetky dokumenty patriace k jednému projektu sú usporiadané do jednej zbierky. Napr. zbierky: Koháry korešpondencia, Collectanea Laučeka, Hurban listy, Abrahamides kázne, Reambulačné protokoly, Rukopisné katalógy, Andrej Kmeť korešpondencia a pod. Jedna zbierka môže obsahovať viac dokumentov. A dokumenty pozostávajú z jednej alebo viacerých strán dokumentu. Každá zbierka v Transkribe má jedinečný identifikátor (ID). Každý dokument v zbierke má jedinečný číselný identifikátor, názov dokumentu, počet strán dokumentu, meno osoby, ktorá nahrala dokument do Transkribu, dátum a čas nahratia, vlastníka zbierky. V zbierke je možné manažovať – tvoriť, vymazať, upravovať, pridávať a upravovať oprávnenie používateľov zbierky so súhlasom a rozhodnutím vlastníka zbierky,

pracovať s kreditmi k zbierke. Ku každému dokumentu je možné popísať všeobecné metadáta a metadáta ku jednotlivej strane, ako aj štrukturálne a textové metadáta a komentáre. Používateľ môže mať niekoľko zbierok s rôznymi dokumentmi. Na účely prezentačnej vrstvy *read&search* je potrebné vytvoriť jednu spoločnú zbierku. Všetky zbierky a dokumenty v Transkribe sú súkromné.

ISBN 978-80-557-2020-3