MATERIÁLY 87-108

O možnostiach automatického prepisu historických rukopisných dokumentov pomocou platformy Transkribus

OTO TOMEČEK

Tomeček, Oto: On the Possibilities of Automatic Transcription of Historical Handwritten Documents Using the Transkribus Platform

The study deals with the problem of automatic transcription of historical handwritten documents. In the first part, the author draws attention to the Transkribus tool, which is one of the platforms that can be used for the purpose of automatic transcription of manuscripts. In the second part, the author presents his personal experience of working with this tool. Using the example of working with a manuscript, the Banská Bystrica town's reambulatory protocol from 1820, he presents the steps of the work, starting with the import of the document, its segmentation, the training of the custom models, up to the automatic transcription and export of the transcribed document. The model trained on the transcription of the reambulatory protocol, together with the Slovak Supermodel trained on several manuscripts, is finally applied by the author to the transcription of another manuscript version of the same protocol. Finally, the author presents the results of the above-mentioned experiments and reflects on the effectiveness of the automatic transcription tool in the work of a historian.

Key Words *Digital Humanities; Transkribus; Handwritten Document; Automatic Transcription; Reambulatory Protocol*

Contact *Oto Tomeček Univerzita Mateja Bela v Banskej Bystrici; oto.tomecek@umb.sk*

DOI – *10.15452/Historica.2025.16.0005*

Úvod

Predmetná štúdia sa zaoberá aktuálnou otázkou využitia umelej inteligencie v historikovej práci. Jej primárnym cieľom je predstaviť relatívne nový nástroj umelej inteligencie – platformu *Transkribus*, predovšetkým však podeliť sa o osobné skúsenosti s prácou na automatickej transkripcii vybraných historických rukopisných textov realizovaných prostredníctvom nej.[1] Na príklade vlastného experimentu, automatického prepisu dvoch rôznych rukopisných exemplárov reambulačného protokolu Banskej Bystrice z roku 1820, poukážem na možnosti a limity tohto nástroja, aby si každý záujemca o prácu s ním urobil vlastný obraz o jeho efektivite.

Umelá inteligencia a možnosti jej praktického uplatnenia patria k intenzívne pertraktovaným témam súčasného sveta. Čím ďalej tým viac umelá inteligencia zasahuje do každodenného života, pričom za jej najčastejšie využívané nástroje možno považovať

[1] Táto štúdia vznikla ako výstup projektu podporeného 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*.

rôzne, dnes už bežne dostupné, jazykové prekladače a generátory textu (chatboty). Využívanie moderných digitálnych technológií, vrátane umelej inteligencie, vo svojom výskume začali v poslednom období intenzívnejšie uplatňovať aj bádatelia v oblasti humanitných vied. V súvislosti s tým tak môžeme sledovať cielený rozvoj tzv. digitálnej humanistiky či digitálnych humanitných vied (digital humanities).[2] Jej základom je pokročilá digitalizácia, ktorá presahuje doterajšie vnímanie digitalizácie ako nástroja určeného pre potreby zachovania kultúrneho dedičstva, vrátane historických písomností, výlučne prostredníctvom jeho prevodu z analógovej do digitálnej podoby. Pod digitalizáciou a výstupmi z oblasti digital humanities dnes už rozumieme predovšetkým vytváranie rozličných elektronických databáz, mapových výstupov či komplexných pamäťových portálov s navzájom prepojenými informáciami rôzneho druhu a obsahu. Okrem digital humanities, ktoré môžeme chápať širšie ako prienik medzi svetom technológií a humanitných vied, dnes už pracujeme aj s užším pojmom digitálna história (digital history), ktorý definuje prienik medzi historickými vedami a modernými technológiami.[3]

Základom a akýmsi odrazovým mostíkom tvorivej práce historika naďalej zostáva hlavne analýza dobového písomného prameňa. Práca s historickými, prevažne rukopisnými, dokumentmi však predstavuje pomerne zdĺhavý a stereotypný proces. Historický dokument je potrebné v archíve identifikovať, dešifrovať, prečítať, následne prepísať a podľa potreby aj preložiť. Moderné technológie vytvorené a zdokonalené v prvej štvrtine 21. storočia umožňujú tieto procesy historikovej práce výrazne zjednodušiť a urýchliť. V online priestore dnes už bežne nachádzame databázy zdigitalizovaných historických dokumentov, ktoré umožňujú vyhľadávať konkrétne dokumenty prostredníctvom prednastavených vyhľadávačov a filtrov.[4] Iné, ktoré nie sú sprístupnené touto formou, dokážeme previesť do digitálnej podoby ich skenovaním alebo odfotením na smartphone priamo v archívoch. Rôzne nástroje a platformy umelej inteligencie umožňujú aj ďalšiu pokročilejšiu prácu s takýmito digitalizátmi. V súčasnosti sú dostupné efektívne nástroje umelej inteligencie umožňujúce ich prepis,[5] rovnako tak ich preklad.[6] Jedným z takýchto nástrojov je aj veľmi dynamicky sa rozvíjajúca platforma *Transkribus*, primárne určená na automatický prepis historických rukopisných textov.

---

[2] Problematike sa v poslednom období venovalo viacero prác. Spomedzi nich možno uviesť aspoň niekoľko príkladov: JANNIDIS, Fotis – KOHLE, Hubertus – REHBEIN, Malte: *Digital Humanities : Eine Einführung*. Stuttgart 2017; FOLTÝN, Tomáš: Digital Humanities : stručné shrnutí stávajícího stavu problematiky v ČR. *ITlib : Informačné Technológie a Knižnice* 21, 2017, č. 3, s. 19–22; BURDICKOVÁ, Anne a kol.: *Digital Humanities*. Praha 2019; GOGORA, Andrej: Ako a prečo pomenovať to, čo robíme? : Problém slovenského prekladu termínu „digital humanities". *Slovenská literatúra* 67, 2020, č. 6, s. 598–613; KATUŠČÁK, Dušan: Digital humanities a automatická transkripcia rukopisných textov. *ITlib : Informačné Technológie a Knižnice* 24, 2020, č. 1, s. 6–16.

[3] MILLIGAN, Ian: *The Transformation of Historical Research in the Digital Age*. Cambridge – New York – Melbourne – New Delhi – Singapore 2022, s. 7.

[4] Väčšina spomedzi týchto databáz sprístupňuje len nasnímané digitalizáty originálnych dokumentov. Do tejto skupiny môžeme zaradiť napríklad často používané voľne dostupné databázy: *Europeana*, *Hungaricana*, *Slovakiana* či *Pammap*. Zatiaľ len málo databáz sprístupňuje aj transkribované dokumenty. Medzi takéto môžeme zaradiť napríklad databázu *Early Modern Manuscripts Online* (EMMO), ktorá sprístupňuje transkribované rukopisy zo 16. a 17. storočia, alebo digitálnu knižnicu *HathiTrust* obsahujúcu množstvo historických kníh a dokumentov, z ktorých niektoré boli zdigitalizované a transkribované pomocou OCR technológie.

[5] Automatickú transkripciu historických rukopisných dokumentov umožňujú napríklad aplikácie: eScriptorium, OCR4all, PERO – OCR a iné.

[6] Medzi najznámejšie prekladateľské softvéry v súčasnosti patria napríklad: Google Translate, DeepL, Wordvice AI či Mirai Translate. Ako najnádejnejší nástroj využiteľný aj na preklad historických dokumentov sa v súčasnosti javí aplikácia ChatGPT, ktorej výhodou je možnosť vlastného trénovania.

Platforma *Transkribus*

Transkribus je v súčasnosti jednou z najpokročilejších platforiem pre automatický prepis historických písomných prameňov. Ide o nástroj umelej inteligencie, cielene vytvorený pre potreby efektívnej práce s historickými dokumentmi rukopisnej aj tlačenej povahy. Nástroj je založený na báze strojového učenia, pomocou ktorého dokáže rozpoznávať štruktúru textu, priradiť k jednotlivým znakom dokumentu iné zodpovedajúce znaky a následne na základe toho zrealizovať automatický prepis textu dokumentu.[7] Keďže *Transkribus* pracuje na báze stotožnenia znakov obsiahnutých v dokumente s požadovanými zodpovedajúcimi znakmi, umožňuje transkripciu rukopisného aj tlačeného dokumentu bez ohľadu na typ písma, použitý jazyk, ako aj dobu a spôsob jeho vyhotovenia.

Platforma *Transkribus* bola vyvinutá konzorciom pod vedením Güntera Mühlbergera z Univerzity v Innsbrucku (Universität Innsbruck)[8] v rámci riešenia projektu Horizont 2020 READ (Recognition and Enrichment of Archival Documents). V roku 2019 sa konzorcium READ zmenilo na komerčnú spoločnosť READ-COOP, ktorá sa stala základňou pre udržanie a ďalší rozvoj poskytovaných služieb a nástrojov platformy. V súčasnosti má spoločnosť viac ako 200 členov z 30 rôznych krajín sveta. Jediným univerzitným pracoviskom v rámci krajín stredovýchodnej Európy, ktoré patrí medzi členov spoločnosti, je Univerzita Mateja Bela v Banskej Bystrici. Okrem nej má spomedzi krajín tohto regiónu zastúpenie medzi členmi spoločnosti už len Krajinská Széčéniho knižnica (Országos Széchényi Könyvtár) v Budapešti.[9] Jednotliví členovia spoločnosti sa podieľajú na ďalšom vývoji a zdokonaľovaní platformy. READ-COOP sa dnes profiluje ako spoločnosť, ktorá prostredníctvom nástroja *Transkribus* umožňuje poskytovanie komplexného súboru služieb založených na prepise, rozpoznávaní a vyhľadávaní historických dokumentov pre vedcov, študentov, knižnice, archívy, ale aj amatérskych historikov. Aktuálne má platforma viac ako 150 000 registrovaných používateľov a prostredníctvom nej bolo doposiaľ spracovaných viac ako 50 000 000 strán dokumentov. Na platforme bolo vytrénovaných viac ako 20 000 modelov určených na automatický prepis dokumentov, spomedzi ktorých je viac ako 150 verejne prístupných.[10]

---

7 *What is Transkribus?* Online, cit. 20. 1. 2025, dostupné na https://readcoop.eu/Transkribus/.

8 Günter Mühlberger pracuje na katedre (oddelení) nemeckého jazyka a literatúry. Od polovice 90. rokov 20. storočia sa vo svojom výskume zameriava na digitalizáciu, digital humanities, digitálne knižnice a problematiku automatického rozpoznávania textu. Od roku 2016 bol hlavným koordinátorom projektu Horizont 2020 READ. V súčasnosti je predsedom rady riaditeľov spoločnosti READ-COOP.

9 Hoci Česká republika nemá inštitucionálne zastúpenie medzi členmi spoločnosti, platforma je známa a používaná aj viacerými tunajšími výskumníkmi. Na platforme je v súčasnosti prístupný model *Old Czech Handwriting* vytrénovaný na takmer 98 000 slovách. Tento model, ktorého autormi sú Anna Michalcová, Martina Spěváčková, Martina Kramarič, Jozef Mikloš, Julie Mizerová, Daniel Katscher, Jitka Filipová, Eva Pasáčková, Martina Vassmar, Eliška Pěnkavová, Leon Glaser a Klára Vučičová, je možné použiť hlavne na prepis rukopisov napísaných v 14. a 15. storočí v gotickej bastarde. Na takmer 150 000 slovách nemeckých a českých rukopisov od 16. do 19. storočia vytrénoval model *Moravian Land Records* Johannes Georg Schwarz. K dispozícii je už aj veľký agregovaný model *Transkribus Czech Handwriting M1* vytrénovaný na 917 580 slovách. Okrem toho je na platforme voľne dostupný aj model určený na prepis českých a slovenských tlačí (*Czech, Slovak Print model M1*) vytrénovaný na takmer 96 000 slovách.

10 Uvedené číselné údaje sa viažu ku dňu 15. 10. 2024. O dynamike, s akou sa spoločnosť rozvíja, svedčí fakt, že na začiatku toho istého roku k 1. 1. 2024 mala spoločnosť ešte len 135 členov, viac ako 100 000 registrovaných používateľov, spracovaných viac ako 40 000 000 strán dokumentov, viac ako 15 000 vytrénovaných modelov, z ktorých bolo viac ako 100 verejne prístupných. *What is Transkribus?* Online, cit. 3. 1. 2024, dostupné na https://readcoop.eu/Transkribus/.

Ako návod na prácu v platforme *Transkribus* slúži niekoľko príručiek a manuálov. Jedna z nich je dostupná online na internetovej stránke spoločnosti READ-COOP.[11] K dispozícii je však aj niekoľko ďalších online manuálov,[12] ako aj webinárov oboznamujúcich s prostredím platformy.[13] Najnovšie sú k dispozícii už aj dve po slovensky písané metodické príručky, ktoré vznikli ako študijné pomôcky pre účastníkov dvoch workshopov organizovaných Univerzitou Mateja Bela v Banskej Bystrici v rokoch 2023 a 2024. Prvá sa zameriava na prácu s pôvodnou verziou platformy *Transkribus Expert Client*, ktorá je dostupná len po stiahnutí a inštalácii priamo v osobnom počítači.[14] Druhá príručka slúži ako manuál pre prácu s online verziou *Transkribus app*.[15] Okrem nich prácu s platformou *Transkribus* pri vytváraní vlastných modelov a prepisov predstavujú, na príklade nimi spracovávaných slovacikálnych dokumentov, napokon aj jednotliví riešitelia projektu Skriptor.[16]

V súčasnosti platforma *Transkribus* ešte stále funguje v dvoch verziách. Prvá a pôvodná verzia *Transkribus Expert Client* (v súčasnosti je k dispozícii už aktualizovaná verzia 1. 27. 0 zo 4. decembra 2023) slúžila na jej vývoj a zatiaľ stále ponúka viac pracovných nástrojov v porovnaní s jej druhou verziou. Táto verzia by už nemala byť v budúcnosti aktualizovaná a ďalej rozvíjaná. Svoju ďalšiu budúcnosť spoločnosť READ-COOP spája už len s druhou verziou platformy. Táto druhá verzia označovaná ako *Transkribus app* (pôvodne *Transkribus Lite*) je dostupná online, pričom používateľom ponúka zjednodušené pracovné prostredie bez potreby jej inštalácie v počítači. V porovnaní s pôvodnou verziou je jej výhodou predovšetkým jednoduchá dostupnosť obmedzená výlučne na internetové pripojenie.

Všetky pracovné súbory (digitalizáty písomných dokumentov) sú uložené na serveroch platformy. Prístup k nim je zabezpečený prostredníctvom režimu vzdialeného prístupu, ku ktorému je potrebné len vysokorýchlostné pripojenie k internetu. Tu sa priebežne ukladajú aj všetky úkony, ktoré realizujeme, jednotlivé verzie modelov, ako aj pracovné a finálne prepisy dokumentov.

Postup a etapy práce v systéme *Transkribus*

Práca s dokumentom na platforme *Transkribus* pozostáva z niekoľkých fáz. Po bezplatnom stiahnutí pôvodnej aplikácie, aktuálne po registrácii na webovej platforme, získava každý

---

[11] *How-to Guides*. Online, cit. 20. 1. 2025, dostupné na https://readcoop.eu/Transkribus/resources/how-to-guides/.

[12] *Transkribus Expert Client Einführung : Transkribus Expert Client Kurs* [1–10]. Online, cit. 3. 1. 2024, dostupné na https://www.youtube.com/watch?v=KLtLg5Dui80 & list=PL7UbQtd4qlhI3VJck98N4kDAeuZfy--H0 alebo tiež: *Introducing the New Transkribus*. Online, cit. 3. 1. 2024, dostupné na https://www.youtube.com/watch?v=3mv7_pYDK5E.

[13] Napr.: *Introduction to Transkribus and the New User Interface : Webinar ENGLISH*. Online, cit. 3. 1. 2024, dostupné na https://www.youtube.com/watch?v=LdBYSDMSeC8.

[14] KATUŠČÁK, Dušan – NAGY, Imrich (eds.): *Automatická transkripcia historických dokumentov : metodická príručka na prácu s platformou Transkribus*. Banská Bystrica 2023.

[15] NAGY, Imrich – KATUŠČÁK, Dušan (eds.): *Transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus : metodická príručka pre účastníkov workshopu*. Banská Bystrica 2024.

[16] NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. *Slovenská archivistika* 51, 2021, č. 2, s. 53–67; MALINIAK, Pavol – NAGY, Imrich (eds.): *Digital humanities : Nástroje sprístupňovania historického dedičstva*. Banská Bystrica 2022; KATUŠČÁK, Dušan – NAGY, Imrich (eds.): *Automatická transkripcia slovacikálnych historických dokumentov.* Banská Bystrica 2023.

užívateľ k dispozícii voľné kredity. Pôvodne to bolo 500 voľných kreditov, v súčasnosti je to 100 kreditov na mesiac. Tie neskôr bádateľ využije pri realizovaní automatického prepisu svojho dokumentu. Ďalšie kredity si bádateľ v prípade potreby dokupuje. Ostatné etapy práce na platforme vrátane vytvárania modelov boli doposiaľ bezplatné a nevyžadovali spotrebu kreditov. Aktuálne došlo tiež k čiastočnému spoplatneniu procesu segmentácie dokumentu. Politika spoločnosti ohľadom výšky spoplatnenia uvedených úkonov sa mení. Pred vykonaním každého spoplatneného úkonu je však používateľ vždy dopredu informovaný.[17]

Po prihlásení a vytvorení vlastnej zbierky v prostredí platformy je potrebné nahrať zdigitalizovaný dokument do zvolenej zbierky. Platforma *Transkribus* počíta okrem klasického skenovania dokumentov aj s možnosťou ich núdzového nasnímania prostredníctvom smartphonov. Práve na tento účel bola vyvinutá pracovná pomôcka *ScanTent* a softvér *DocScan app*. Prenosný *ScanTent* (skenovací stan) disponuje vlastným podsvietením, ktoré zabezpečuje rovnomerné osvetlenie dokumentu a eliminuje jeho možné zatienenie počas procesu nasnímania. Obsahuje aj 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, resp. len mierne presahujúcej tento formát. Softvér *DocScan app*, ktorý je možné voľne stiahnuť do mobilného smartphonu, umožňuje zosnímanie dokumentu aj bez neustáleho manuálneho stláčania spúšte fotoaparátu. Bádateľ sa takto môže sústrediť na prácu s dokumentom a otáčanie jeho strán, prípadne výmenu voľných listov. Uvedená aplikácia umožňuje zároveň priamy prenos nasnímaných dokumentov rovno do prostredia platformy.[18]

V prípade, že bádateľ potrebuje nasnímaný dokument upraviť (orezať, zaostriť, zosvetliť a pod.), možno odporučiť stiahnutie dokumentu do osobného počítača a jeho dodatočnú úpravu. Takto upravený dokument možno neskôr nahrať z osobného počítača priamo do zvolenej zbierky v prostredí platformy *Transkribus*. Import digitalizátov a následná práca v expert klientovi je možná len so súbormi podporovaných formátov PDF, JPEG, PNG a TIFF. V rámci webovej aplikácie vypadla možnosť vkladania súborov TIFF. Digitalizáty jednotlivých strán dokumentu vo formátoch JPEG a PNG by nemali presahovať veľkosť 10 MB. Rozlíšenie snímok by malo ideálne dosahovať hodnotu 300 dpi, avšak z vlastnej skúsenosti možno potvrdiť, že *Transkribus* dokáže pracovať aj so snímkami vyhotovenými v podstatne nižšom rozlíšení. Vyššie rozlíšenie je zbytočné, lebo nijako neprispieva k zlepšeniu rozpoznania a automatickej transkripcie textu. V prípade vkladania dokumentu vo formáte PDF by nemala jeho veľkosť presahovať 200 MB, pričom maximálny počet vložených strán v rámci jedného súboru by nemal presahovať hodnotu 3 000.

V rámci dnes už nepodporovanej platformy expert klient je možné využiť aj iné spôsoby vkladania digitalizátov. Prvým spôsobom je cesta *Private FTP*, ktorá umožňuje nahrať viac súborov/priečinkov naraz. Tento spôsob si vyžaduje inštaláciu klienta *FTP*. Pri druhom spôsobe nahrávania dokumentov, cestou *IIIF Manifest*, sa dokumenty vkladajú priamo z webovej stránky po skopírovaní a vložení URL adresy. Tento spôsob je možné použiť

[17] Aktuálne si bádateľ môže k voľne získaným kreditom dokúpiť 1 000 kreditov za sumu 238,80 €, pričom spotreba kreditov je nasledovná: za prepis jednej strany rukopisu sa odráta 1 kredit a za prepis jednej strany tlačeného textu sa odráta 0.5 kreditu. Segmentácia jednej strany textu s riadkami má aktuálne hodnotu 0.25 kreditu. Pri segmentácii strany s tabuľkovou formou bude odrátaný 1 kredit. Uvedená cenová politika je platná ku dňu 15. 10. 2024.

[18] K obom uvedeným pomôckam pozri: *Scan Documents on the Go with the ScanTent*. Online, cit. 20. 1. 2025, dostupné na https://readcoop.eu/scantent/.

vtedy, ak inštitúcia (archív, knižnica) poskytuje online prístup k svojim zdigitalizovaným dokumentom prostredníctvom štandardu *IIIF*. Aj pri treťom spôsobe nahrania dokumentov, cestou *DFG Viewer METS*, sa dokumenty vkladajú na platformu priamo z webu jednoduchým vložením URL adresy do určeného poľa. Do budúcnosti sa očakáva, že všetky tieto spôsoby budú dostupné aj na webovej aplikácii platformy.[19]

Po nahraní dokumentu do prostredia platformy nasleduje najzdĺhavejší proces segmentácie dokumentu. Pod pojmom segmentácia rozumieme úpravu dokumentu do tej podoby, aby bol pripravený na realizovanie samotnej transkripcie. Počas segmentácie musíme vyznačiť textové polia dokumentu (*Text Regions*), šírku jednotlivých riadkov (*Line Regions*) a základné čiary týchto riadkov (*Baselines*). Pre lepšiu orientáciu je každý z týchto prvkov označený iným prednastaveným podsvietením. Súčasťou procesu segmentácie je aj zadefinovanie správneho poradia čítania textových polí a jednotlivých riadkov v rámci textových polí. V prípade vnútorne nečleneného textu predstavuje každé textové pole zvyčajne jednu stranu. V prípade vnútorne členených textov (napr. vo forme tabuliek) je potrebné text na jednej strane rozdeliť do viacerých textových polí. Čítanie jednotlivých riadkov je potrebné nastaviť individuálne podľa jednotlivých textových polí.

Segmentáciu môžeme realizovať automaticky alebo manuálne. Automatickú segmentáciu za nás realizuje samotný softvér. Hoci automatická segmentácia dokumentu prebehne pomerne rýchlo, vyžaduje si pomerne zdĺhavú kontrolu a prípadnú korekciu chýb. Manuálna segmentácia je časovo náročnejšia, keďže pozostáva z manuálneho nastavenia textových polí. Nastavenie riadkov a základných čiar prebieha aj v tomto prípade automaticky, preto je aj v tomto prípade nevyhnutná dodatočná kontrola a prípadná korekcia chýb.[20]

Súčasťou korekcie chýb pri automatickej aj manuálnej segmentácii je aj kontrola dĺžky základných čiar riadkov, prípadne aj úprava šírky riadkov. Pri procese segmentácie sa pomerne často stáva, že základné čiary nesiahajú presne od prvého po posledný znak v riadku. V takomto prípade je potrebné základnú čiaru ručne natiahnuť na dĺžku celého riadku. V prípadoch, že niektoré znaky presahujú z jedného riadku do druhého riadku, je vhodné upraviť na inkriminovaných miestach aj šírku riadku. Toto umožňujú body nachádzajúce sa na okrajoch podsvietenia riadku. Tieto body je možné manuálne ťahať rôznymi smermi a vytvárať tak polygóny, ktoré sa prispôsobujú tvaru jednotlivých písmen v riadku.

Po ukončení segmentácie možno pristúpiť k samotnej transkripcii dokumentu. Pred automatickou transkripciou dokumentu je potrebné vybrať vhodný model, ktorý umožní tento prepis zrealizovať. Bádateľ si môže vybrať z voľne dostupných modelov. Špecifiká dokumentu, predovšetkým osobitý rukopis jeho pisateľa, však väčšinou neumožňujú vyberať z dostupnej a v súčasnosti zatiaľ aj pomerne obmedzenej ponuky modelov. Z uvedeného dôvodu je nevyhnutné vytvoriť pre potreby automatickej transkripcie vlastný model ušitý na mieru konkrétnemu dokumentu. Možnosť vytvorenia vlastného modelu a jeho ďalšie trénovanie možno považovať za pridanú hodnotu platformy *Transkribus*.[21]

Pri vytváraní vlastného modelu je potrebné najprv realizovať manuálny prepis vybranej

[19] KATUŠČÁK, D. – NAGY, I. (eds.): Automatická transkripcia, s. 38; resp. TÍŽ: Transkripcia, s. 44–45.

[20] Postup prác pri obidvoch spôsoboch segmentácie je podrobne rozpísaný v metodických príručkách pripravených riešiteľmi projektu Skriptor. Pozri: KATUŠČÁK, D. – NAGY, I. (eds.): Automatická transkripcia, s. 40–84; TÍŽ: Transkripcia, s. 46–84.

[21] KATUŠČÁK, D.: Digital humanities, s. 9.

časti dokumentu. Manuálny prepis je potrebné urobiť formou transliterácie, teda každé písmeno prepísať presne tak, ako je uvedené v dokumente. Dôležité je neopravovať žiadne pisárske chyby ani nerozpisovať použité skratky. Na základe manuálneho prepisu naučíme softvér priradiť každému rukopisnému znaku konkrétny požadovaný znak. Tvorcovia platformy *Transkribus* odporúčajú manuálne prepísať text v rozsahu minimálne 10 000 slov, ideálne dokonca až 15 000 slov.[22]

Po ukončení manuálneho prepisu vybranej vzorky požadovaného rozsahu je potrebné túto vzorku rozdeliť na dve časti, ideálne v pomere 10:1. Rozsiahlejšia časť sa použije ako vzorka pre tréningový (cvičný) set, na ktorej sa stroj učí rozpoznávať jednotlivé znaky a priraďovať ich k požadovaným znakom. Menšia časť textu, predstavujúca zhruba jednu desatinu manuálne prepísanej vzorky dokumentu, sa použije ako vzorka pre validačný (overovací) set, na príklade ktorej softvér zrealizuje automatickú transkripciu vychádzajúc z toho, čo sa sám naučil na vzorke zaradenej do tréningového setu. Pri tvorbe modelu bolo možné ešte donedávna voliť spomedzi dvoch možností strojového učenia, ktorými boli nástroje *HTR+* a *PyLaia*. V súčasnosti je možné realizovať automatickú transkripciu už len prostredníctvom nástroja *PyLaia*.

Smerodajným ukazovateľom pri vyhodnotení úspešnosti modelu je údaj o chybovosti vo validačnom sete. *Transkribus* dokáže vyhodnotiť chybovosť na úrovni slov (WER, t. j. *Word Error Rate*), ako aj na úrovni znakov (CER, t. j. *Character Error Rate*). Chybovosť na úrovni slov (WER) by nemala presahovať 30 % a na úrovni znakov (CER) 15 %, ideálne dokonca 10 %.[23] Pri dosiahnutí týchto a nižších hodnôt je transkribovaný text možné považovať za zrozumiteľný a vhodný na ďalšie použitie. Ideálne by chybovosť na úrovni slov aj znakov mala dosahovať čo najnižšie hodnoty. Výsledok pod 5 % chybovosti na úrovni znakov sa už považuje za vynikajúci výsledok. Nižšiu ako 2 % chybovosť na úrovni znakov vo validačnom sete je možné dosiahnuť zvyčajne len pri tlačených textoch.

Vlastný model je možné v nasledujúcich fázach trénovaním postupne vylepšovať až do najnižšej možnej chybovosti vo validačnom sete. Trénovanie a vylepšovanie modelu môže prebiehať dvoma spôsobmi. Prvým spôsobom je neustále rozširovanie manuálne prepísanej vzorky textu a postupné rozširovanie počtu slov zaradených do tréningového setu. Druhým spôsobom je využitie dostupného *Base modelu* (základného modelu) pri trénovaní vlastného modelu. Za *Base model* možno považovať skôr vytrénovaný a v platforme *Transkribus* voľne dostupný model. Ako *Base model* možno použiť aj vlastný model vytvorený počas skoršieho trénovania predmetného dokumentu alebo iný voľne prístupný model vytrénovaný iným bádateľom na type písma najviac zodpovedajúcemu písmu dokumentu, ktorý je predmetom transkripcie. Vytrénovanie modelu na úrovni chybovosti znakov vo validačnom sete od 2 do 5 % znamená de facto úspešnosť prepisu na úrovni znakov dosahujúcu hodnoty 95–98 %. Vytrénovanie modelu s takýmito parametrami možno považovať za excelentný výsledok.

V tejto súvislosti je potrebné upozorniť, že model s najnižším percentom chybovosti vo validačnom sete nemusí byť vždy najvhodnejším modelom pre finálnu automatickú transkripciu vybraného dokumentu. Na výslednú chybovosť validačného modelu vplýva

---

[22] V prípade tlačeného dokumentu by mal postačovať prepis textu obsahujúci 5 000 slov. K uvedeným hodnotám bližšie porovnaj: MUEHLBERGER, Guenter et al.: Transforming Scholarship in the Archives through Handwritten Text Recognition : Transkribus as a Case Study. *Journal of Documentation* 75, 2019, no. 5, s. 959; NAGY, I.: Možnosti, s. 56; KATUŠČÁK, D. – NAGY, I. (eds.): Automatická transkripcia, s. 91.

[23] KATUŠČÁK, D.: Digital humanities, s. 14; NAGY, I.: Možnosti, s. 59–60.

niekoľko faktorov. Okrem počtu slov zaradených do tréningového setu je to hlavne výber strán zaradených do tréningového a validačného setu, ako aj výber *Base modelu* použitého pri tvorbe vlastného modelu. Vo všeobecnosti platí, že je vhodnejšie použiť na automatickú transkripciu model vytrénovaný na viacerých, minimálne 10 000, slovách. V prípade, že do tréningového a validačného setu zaraďujeme strany s rukopisným textom nasnímaným v menšej kvalite, sa tento fakt môže negatívne premietnuť do výslednej chybovosti modelu. Ak naopak volíme strany nasnímané vo vyššej kvalite, chybovosť sa zákonite znižuje. Chybovosť modelu môže znižovať aj zaradenie strán s čitateľnejším rukopisom, s menším počtom škrtov alebo iných pisárskych omylov. Dôležitý je zvyčajne aj výber a zaradenie *Base modelu* pri trénovaní. Vhodný výber *Base modelu* zvyčajne pomáha výrazne znižovať chybovosť vlastného modelu.[24]

Po vytrénovaní a zvolení najvhodnejšieho modelu možno pristúpiť k samotnej automatickej transkripcii zvyšného, doposiaľ neprepísaného, textu dokumentu. Po ukončení automatickej transkripcie nasleduje kontrola prepísaného textu a korekcia prípadných chýb. Prepísaný text možno pri kontrole doplniť tagmi, ktoré môžu v budúcnosti uľahčiť ďalšiu prácu s dokumentom. *Transkribus* rozlišuje textové a štrukturálne tagy. Textové tagy majú zvyčajne vysvetľujúci význam. Týmto spôsobom môžeme označovať napríklad osoby, inštitúcie, geografické názvy, dátumy, skratky alebo nečitateľné slová v predmetnom texte. Štrukturálne tagy definujú štruktúru dokumentu, takže pomocou nich môžeme označovať napríklad nadpisy, odseky, čísla strán alebo marginálie.[25]

Po ukončení niektorej zo zásadných etáp úpravy dokumentu v platforme *Transkribus* je možné vždy meniť označenie statusu, resp. stavu úpravy každej strany dokumentu. Pri tomto procese je možné vyberať spomedzi možností *New*, *In Progress*, *Done*, *Final*, *Ground Truth*. Každý status je v celkovom prehľade (*Overview*) odlíšený farebne. To umožňuje rýchlu orientáciu a rozpoznanie stavu spracovania dokumentu. Finálnu verziu transkribovaného a skontrolovaného dokumentu označuje status *Ground Truth*, ktorý je nevyhnutný pri trénovaní modelu automatickej transkripcie.

Poslednou fázou práce v *Transkribe* je exportovanie prepísaného dokumentu. To je možné zrealizovať dvoma spôsobmi. V prípade, že zvolíme možnosť exportu zo servera (*Server export*), dokument sa uloží na serveri, odkiaľ sa stiahne prostredníctvom priameho odkazu na link servera. Pri druhej možnosti exportu z klienta (*Client export*) môžeme dokument uložiť priamo do osobného počítača. Po výbere spôsobu exportu zvolíme formát exportovaného dokumentu. Aj v tomto prípade je možné vyberať z viacerých možností, napríklad *Transkribus* Dokument, PDF, TEI, DOCX alebo EXCEL. Pridanou hodnotou je možnosť exportovať dokument aj podľa tagov. V tomto prípade môžu byť tagy v dokumente osobitne zvýraznené alebo je možné vytvoriť osobitný súbor pre ľubovoľnú kategóriu tagov.

Vyexportovaný a uložený dokument významne uľahčuje ďalšiu prácu s dokumentom podľa záujmu bádateľa. Zdigitalizovaný dokument prepísaný automatickou transkripciou môže byť podkladom neskoršieho vydania dokumentu ako pramennej edície. Dokument však môže slúžiť aj na jednoduché vyhľadávanie konkrétnych údajov z dokumentu prostred-

[24] Tento poznatok sa potvrdil napríklad pri vytváraní modelov automatickej transkripcie určených na prepis kanonických vizitácií. Porovnaj: KATRENIAK, Martin: *Automatická transkripcia rukopisných historických textov na príklade vybraných kanonických vizitácií*. Diplomová práca. Banská Bystrica 2022; KATRENIAK, Martin – KUNEC, Patrik: Automatická transkripcia historických prameňov obsahujúcich viac rukopisov na príklade kanonických vizitácií. In: MALINIAK, P. – NAGY, I. (eds.): Digital humanities, s. 49–50.

[25] Bližšie o spôsobe tagovania dokumentu v Transkribe pozri: KATUŠČÁK, D. – NAGY, I. (eds.): Automatická transkripcia, s. 117–128; resp. TÍŽ: Transkripcia, s. 113–120.

níctvom fulltextového vyhľadávača alebo prostredníctvom zvolených tagov. Pozitívom je, že po takomto spracovaní dokumentu možno realizovať fulltextové vyhľadávanie aj v PDF verzii originálneho dokumentu, nielen toho transkribovaného.

Transkripcia vybraného dokumentu reambulačného protokolu A

Na účely realizovania experimentu automatickej transkripcie som si vybral rukopisný dokument reambulačného protokolu mesta Banská Bystrica z roku 1820. Dokument predstavuje podrobný opis obchôdzky mestských hraníc s detailným zaznačením ich priebehu prostredníctvom určených hraničných bodov. Protokol sa zachoval vo veľmi dobrom stave v dvoch obsahovo zhodných exemplároch vyhotovených rozličnými pisármi. V prípade oboch exemplárov dokumentu sa jedná o čistopis napísaný humanistickou (kancelárskou) kurzívou v latinskom jazyku. Prvý exemplár (pre potreby tejto štúdie označený ako reambulačný protokol A), ktorý sa zachoval uložený v komorskom archíve v Banskej Štiavnici,[26] bol napísaný jednou rukou menovite neznámeho pisára. Druhý exemplár (reambulačný protokol B) bol pôvodne uložený v archíve mesta Banská Bystrica a bol celý napísaný rukou iného neznámeho pisára.[27] Pre potreby experimentu automatickej transkripcie som si vybral reambulačný protokol A (banskoštiavnický exemplár). Druhý, reambulačný protokol B (banskobystrický exemplár), neskôr poslúžil pre overenie funkčnosti vytvorených modelov automatickej transkripcie. Pri prepise oboch exemplárov reambulačného protokolu Banskej Bystrice z roku 1820, vyhotovených dvoma rôznymi pisármi, som použil výlučne pôvodnú verziu platformy *Transkribus Expert Client*.

Nasnímanie reambulačného protokolu A (ďalej RPA) prebehlo v študovni Slovenského banského archívu v Banskej Štiavnici pomocou skenovacieho stanu a dvoch smartphonov – Huawei P40 Pro a Google Pixel 4. Ani jeden zo smartphonov nemal stiahnutú aplikáciu *DocScan app*. Prostredníctvom smartphonu Huawei P40 Pro boli vyhotovené snímky dokumentu s rozlíšením 96 dpi a rozmermi 4096 x 3072 pixelov. Pomocou druhého smartphonu bol vyhotovený záložný súbor so snímkami dokumentu s rozlíšením 72 dpi a rozmermi 4 032 x 3 024 pixelov. Celý dokument bol nasnímaný na 129 snímkach, pričom jedna snímka predstavovala dve pôvodné strany dokumentu. Spomedzi 258 nasnímaných pôvodných strán dokumentu sa rukopisný text nachádzal na 245 stranách. Zvyšok predstavovali prázdne strany a vložené mapové listy.

Po importovaní dokumentu do platformy *Transkribus* prebehla segmentácia všetkých jeho strán.[28] Napriek tomu, že prevažnú časť dokumentu predstavovali strany s tabuľkovým členením textu a doplňujúcich údajov,[29] rozhodol som sa zrealizovať automatickú segmentáciu dokumentu, vhodnejšiu skôr pre vnútorne nečlenený text. Na tých stranách dokumentu, ktoré neobsahovali žiadne tabuľky, ale len samotný text, prebehla automatická segmentácia bez väčších problémov. Takýchto strán však obsahuje dokument len minimum.

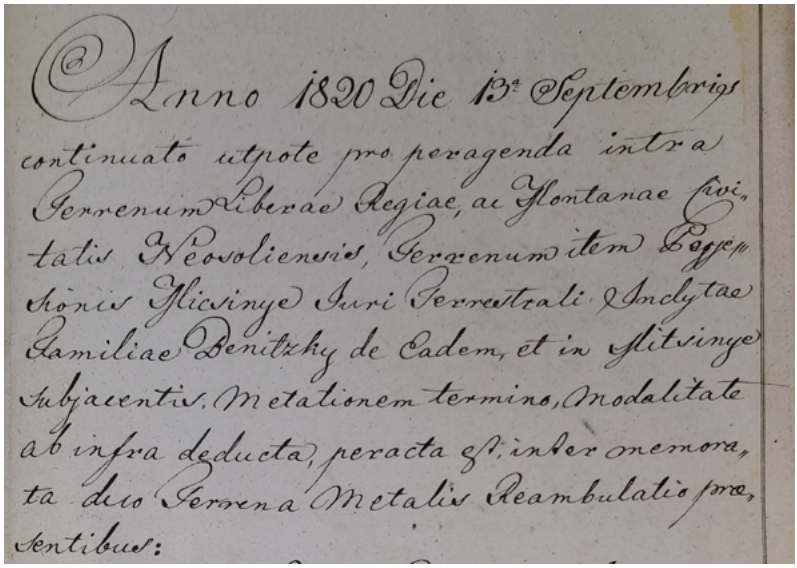
---

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

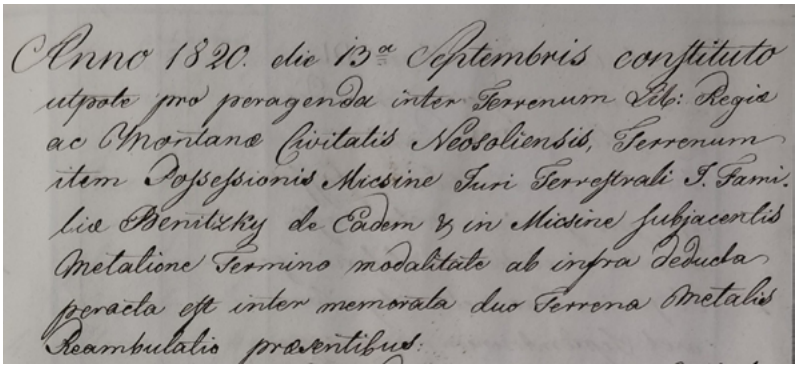
[27] Štátny archív v Banskej Bystrici, fond Mesto Banská Bystrica, príručná knižnica, ev. č. 134.

[28] Proces segmentácie dokumentu podrobne predstavuje práca: TOMEČEK, Oto: Automatická transkripcia reambulačného protokolu Banskej Bystrice z roku 1820. In: KATUŠČÁK, D. – NAGY, I. (eds.): Automatická transkripcia slovacikálnych historických dokumentov, s. 111–114.

[29] Išlo o číselné údaje vyjadrujúce poradie jednotlivých hraničných úsekov, rovnako tak vzdialenosti a uhly medzi jednotlivými hraničnými bodmi.



Obr. 1: Ukážka rukopisu reambulačného protokolu A, autor O. Tomeček



Obr. 2: Ukážka zhodnej časti textu v rukopise reambulačného protokolu B, autor O. Tomeček

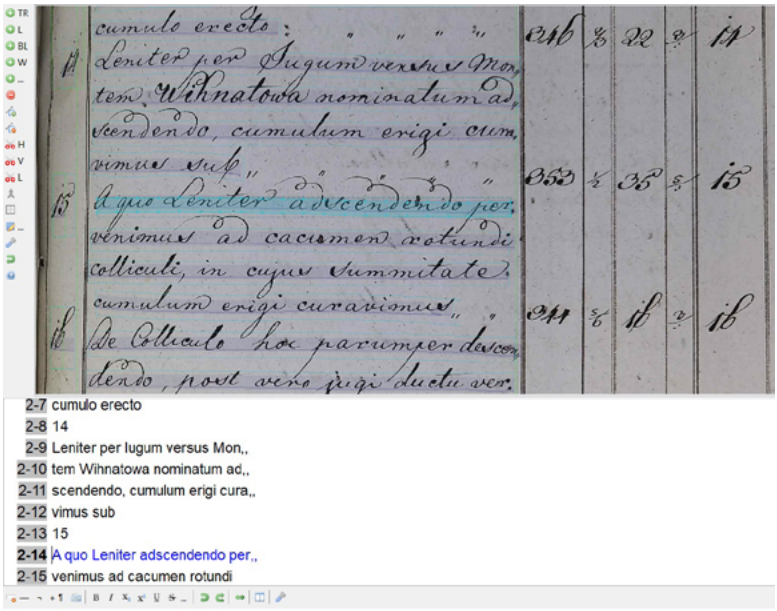
V prípade zvyšných strán, vytvorených tabuľkovou formou, sa už automatická segmentácia ukázala ako nie príliš vhodný nástroj úpravy dokumentu. Po automatickej segmentácii tieto strany obsahovali príliš veľa chýb.

Všetky nepresnosti a chyby vzniknuté pri automatickej segmentácii dokumentu bolo potrebné v nasledovnej fáze korigovať manuálnou úpravou. Najprv boli vymazané všetky nepotrebné znaky označené počas automatickej segmentácie, vrátane nepotrebných číselných údajov, hlavičiek tabuliek, ako aj všetkých omylom vyznačených údajov. Z doplňujúcich číselných údajov bolo ponechané len označenie poradia jednotlivých hraničných úsekov, ktoré ako jediné spomedzi číselných údajov malo väčší význam na pochopenie a správnu interpretáciu obsahu dokumentu. V ďalšej fáze prebehla kontrola poradia čítania textových polí a jednotlivých riadkov a ich prípadná korekcia. Nakoniec prebehla kontrola označenia dĺžky a šírky každého riadku.

Po ukončení segmentácie celého dokumentu mohla začať jeho automatická transkripcia. Tá je však možná len v tom prípade, keď je k dispozícii vhodný model určený na transkripciu. Obmedzená ponuka voľne dostupných modelov automatickej transkripcie bohužiaľ neumožňovala širokú škálu výberu. Spomedzi všetkých som ako najvhodnejší vyhodnotil model *Neolatin Ravenstein 1643–1772*, ktorý vytrénovala Annemieke Romein.[30] Písmo dokumentov, na ktorých bol vytrénovaný tento model, sa najviac približovalo písmu RPA.[31]

Uvedený model som otestoval pri prepise dvoch menších vzoriek predmetného dokumentu RPA. Chybovosť však bola v oboch prípadoch taká vysoká, že neumožňovala zrozumiteľné čítanie a pochopenie textu. Ako nevyhnutná sa preto ukázala potreba vytvoriť vlastný model cielene pripravený na predmetný rukopis dokumentu.

Tvorba vlastného modelu začína manuálnym prepisom vybranej vzorky dokumentu, ktorý má byť predmetom automatickej transkripcie. Spomedzi 245 textových strán celého dokumentu RPA som sa rozhodol v prvej fáze zrealizovať manuálny prepis prvých 49 strán textu (približne jedna pätina z celkového rozsahu textovej časti dokumentu podľa počtu strán a jedna štvrtina z hľadiska počtu slov). Táto vzorka sa neskôr ukázala ako nedostatočná, keďže ju tvorilo len približne 7 200 slov. Po vytvorení prvých dvoch modelov založených na tejto vzorke som v druhej fáze pristúpil k jej rozšíreniu o ďalších desať manuálne



Obr. 3: Ukážka mechanického prepisu RPA v prostredí platformy Transkribus, autor O. Tomeček

[30] U pôvodného modelu vytrénovaného technológiou *HTR+* dosahovala chybovosť na úrovni znakov hodnotu 4,51 % v tréningovom (cvičnom) súbore a 3,58 % vo validačnom (overovacom) súbore. Finálny model (*PyLaia_Neolatin_Ravenstein*) bol vytrénovaný technológiou *PyLaia* na vzorke 64 435 slov dňa 13. 11. 2020. Tento už na úrovni znakov dosahuje chybovosť 1,30 % v tréningovom a 4 % vo validačnom sete.

[31] Dobré skúsenosti s použitím tohto modelu predstavili vo svojej práci aj Martin Katreniak a Patrik Kunec. Pozri: KATRENIAK, M.: Automatická transkripcia... [Diplomová práca], taktiež TÝŽ – KUNEC, P.: Automatická transkripcia, s. 49–50.

prepísaných strán. Spolu som tak zrealizoval manuálny prepis 59 strán textu obsahujúcich 8450 slov. K postupnému rozširovaniu počtu slov zaradených do tréningového a validačného setu pri trénovaní modelov v nasledujúcich fázach dochádzalo potom už len prostredníctvom postupnej automatickej transkripcie ďalších častí textu.

Prvé modely boli vytvorené pomocou dnes už nepodporovaného nástroja strojového učenia označovaného ako *HTR+*.[32] Prvý vlastný model, vychádzajúci z prvej manuálne prepísanej vzorky 49 strán, dosahoval v podstatnom validačnom sete chybovosť na úrovni znakov (CER) 5,35 %. Po použití *Base modelu,*[33] pri trénovaní vlastného modelu, sa podarilo znížiť chybovosť na rovnakej vzorke textu na úroveň 4,74 %.

Po rozšírení manuálne prepísanej vzorky textu o spomenutých desať strán a rozšírení počtu slov zaradených do tréningového setu (zhruba o 2000) bolo možné pristúpiť k vytvoreniu ďalších dvoch modelov automatickej transkripcie. Prvý z nich dosahoval vo validačnom sete 5,09 % chybovosť na úrovni znakov. Po použití rovnakej vzorky textu zaradenej do tréningového a validačného setu a použití *Base modelu* sa podarilo túto chybovosť znížiť na hodnotu 4,25 %.

Všetky ďalšie modely boli vytvorené prostredníctvom nástroja strojového učenia označovaného ako *PyLaia*. Prvé modely vytvorené pomocou tohto nástroja vznikli na príklade rovnakých vzoriek textu ako tie, ktoré boli vytvorené pomocou nástroja *HTR+*. Prvý spomedzi týchto modelov vznikol na pôvodnej vzorke manuálne prepísaného textu, pričom bolo dodržané aj totožné rozdelenie textu medzi tréningový a validačný set. Výsledkom tohto trénovania bol model s chybovosťou 5,90 % na úrovni znakov v dôležitejšom validačnom sete. Pri použití *Base modelu* sa podarilo znížiť chybovosť vlastného modelu vytvoreného na totožnej vzorke textu na 3,51 %. Po použití rozšírenej vzorky textu zaradenej do tréningového setu sa podarilo vytvoriť modely s chybovosťou 5 % bez použitia *Base modelu* a 3,71 % s použitím *Base modelu*.

Uvedené výsledky boli veľmi povzbudivé, keďže model dosahujúci chybovosť pod 5 % na úrovni znakov vo validačnom sete možno vyhodnotiť ako mimoriadne úspešný. Ako opodstatnené a účelné sa ukázalo využívanie základného (*Base*) modelu pri trénovaní vlastného modelu. Zníženie chybovosti modelu, pri rozšírení počtu slov použitých pri jeho trénovaní, zároveň naznačilo možný trend ďalšieho znižovania chybovosti pri rozširovaní vzorky textu v tréningovom sete. Z uvedeného dôvodu som v ďalšej fáze trénovania modelov pokračoval v rozširovaní tejto vzorky, avšak už nie prostredníctvom mechanického prepisu textu, ale postupne pomocou jeho automatickej transkripcie.

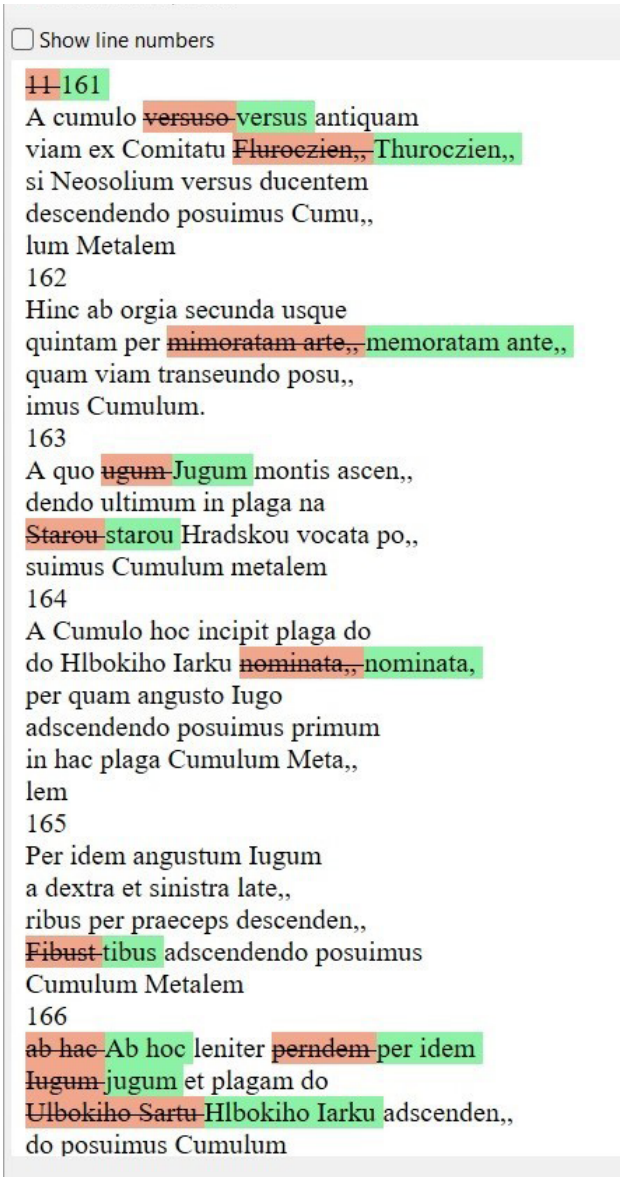
Prvé rozšírenie textu o trinásť strán som realizoval prostredníctvom vlastného modelu (č. 4) dosahujúceho chybovosť 3,71 % na úrovni znakov vo validačnom sete. Po tomto rozšírení prepísanej vzorky bolo potrebné pristúpiť ku korekcii všetkých chýb, ktoré vznikli počas automatickej transkripcie. Takto prepísaný a opravený text mohol byť označený ako *Ground Truth* a následne priradený k ostatnému manuálne prepísanému textu. Po tomto rozšírení prepísaného textu a zaradení necelých 9362 slov do tréningového setu sa podarilo aj s pomocou *Base modelu* vytrénovať nový model (č. 7), ktorý dosahoval chybovosť 2,80 % na úrovni znakov vo validačnom sete.

Pri ďalšom rozšírení prepísanej vzorky textu som využil na automatickú transkripciu práve tento model s chybovosťou 2,80 %. Pomocou rovnakej metodiky a postupného zvyšovania počtu slov (znakov) zaradených do tréningového setu sa postupne podarilo znižo-

[32] Metodika je podrobne rozpísaná v práci: TOMEČEK, O.: Automatická transkripcia, s. 115–120.

[33] Používanie vlastných modelov pre potreby *Base modelu* sa neosvedčilo, preto som vo väčšine prípadov ako *Base model* používal voľne prístupný model *Neolatin Ravenstein* (ďalej NLR).

vať percento chybovosti na úrovni znakov vo validačnom sete najprv na 2,70 % (model 10) a napokon až na 2,60 % (model 11 a 12). Najnižšiu uvedenú chybovosť sa podarilo dosiahnuť na vzorke 12 923 slov zaradených do tréningového setu (model 11). Pri ďalšom rozšírení tréningového setu na 16 943 slov sa percento chybovosti udržalo na rovnakej hodnote 2,60 % (model 12). Na základe uvedeného som usúdil, že ďalším trénovaním sa už percento chybovosti nepodarí výraznejšie znížiť a proces trénovania modelov mohol byť ukončený. Za finálny model určený na automatickú transkripciu zostávajúceho, doposiaľ neprepísaného, textu dokumentu som zvolil druhý z týchto dvoch modelov (č. 12) vytrénovaný na vyššom počte slov.



Obr. 4: Ukážka vyhodnotenia chybovosti automatického prepisu RPA s použitím modelu 12, autor O. Tomeček

Model	Training Set			Validation Set		
	Words	Lines	CER	Words	Lines	CER
Model 1	5 702	1 208	2,70%	1 500	425	5,90%
Model 2 (+ Base m. NLR)	5 702	1 208	0,60%	1 500	425	3,51%
Model 3	7 690	1 773	1,20%	760	233	5,00%
Model 4 (+ Base m. NLR)	7 690	1 773	0,50%	760	233	3,71%
Model 5 (+ Base m. NLR)	7 193	1 790	0,50%	1 257	216	6,90%
Model 6	9 362	2 271	1,20%	711	233	4,40%
Model 7 (+ Base m. NLR)	9 362	2 271	0,70%	711	233	2,80%
Model 8 (+ Base m. NLR)	10 500	2 595	0,70%	997	301	3,41%
Model 9 (+ Base m. M7)	10 500	2 595	0,30%	997	301	3,00%
Model 10 (+ Base m. NLR)	10 735	2 652	0,50%	711	233	2,70%
Model 11_X	12 923	3 288	1,20%	1 210	376	3,51%
Model 11 (+ Base m. NLR)	12 923	3 288	0,50%	1 210	376	2,60%
Model 12 (+ Base m. NLR)	16 943	4 488	0,50%	1 836	583	2,60%

Tab. 1: Vývoj a trénovanie modelov PyLaia určených na prepis RPA

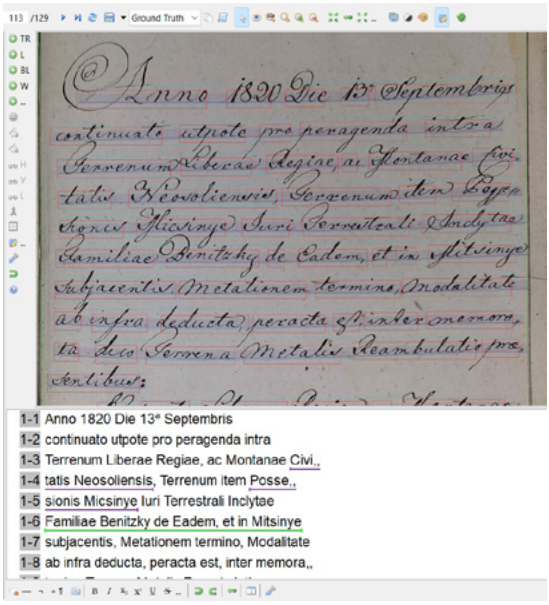
Po automatickom prepise zostávajúcej časti RPA (prostredníctvom modelu 12) bolo potrebné opäť korigovať všetky identifikované chyby. Podľa ich vplyvu na zrozumiteľnosť prepísaného textu ich možno rozdeliť na zanedbateľné a závažné. Medzi zanedbateľné chyby, ktoré neovplyvňujú zrozumiteľnosť prepísaného textu, možno zaradiť nie vždy presné rozlišovanie veľkých a malých písmen, medzier medzi slovami, bodiek na konci vety alebo za radovou číslovkou, resp. čiarok vo vete. Pomerne častou chybou bolo aj vynechanie diakritických znamienok, hlavne mäkčeňa, použitých pri slovensky zapísaných toponymách. Za tento menej závažný typ chyby možno považovať aj nesprávnu identifikáciu rozdeľovníka slov na konci riadku.

Medzi závažnejšie možno zaradiť také chyby, ktoré menia význam a štruktúru slov. Tieto vznikali predovšetkým zámenou niektorých písmen. Na margo väčšiny uvedených chýb je potrebné uviesť, že boli v prevažnej miere spôsobené zaváhaním pisára. Ten zapísal písmeno buď nejednoznačne, alebo často s presahom do iného riadku, kde sa mohlo prekrývať s iným písmenom. Pomerne vysoká chybovosť sa ukázala aj pri identifikácii arabských číslic.[34]

Vzhľadom na fakt, že automatickou transkripciou prepísaný dokument RPA sa stal prvým takto vytvoreným dokumentom na Slovensku, padlo rozhodnutie o jeho vydaní vo forme klasickej pramennej edície. Na základe toho bolo potrebné realizovať dodatočnú kontrolu správnosti celého prepisu.[35] Počas nej som všetky v dokumente uvedené osoby a miestne názvy dodatočne označoval prostredníctvom tagov. Tieto tagy neskôr poslúžili ako základná databáza pre potreby vytvorenia menného a miestneho registra edície.

[34] TOMEČEK, O.: Automatická transkripcia, s. 120.

[35] Pomocnú ruku pri vytváraní finálnych modelov a kontrole správnosti prepisu mi poskytol kolega Imrich Nagy, za čo mu touto cestou vyjadrujem osobitné poďakovanie.



Obr. 5: Ukážka automatického prepisu dokumentu RPA s dodatočným opravami a vloženými tagmi, autor O. Tomeček

Po ukončení kontroly prepisu a tagovania dokumentu prebehol export dokumentov potrebných pre vydanie pramennej edície. Tagy so zoznamom mien osôb a miestnych názvov boli vyexportované ako dva samostatné súbory vo formáte EXCEL. Samotný text dokumentu bol vyexportovaný ako samostatný súbor vo formáte DOCX. Textová časť dokumentu mala podobu hutného textu členeného podľa jednotlivých strán pôvodného dokumentu. Keďže takto vyexportovaný text nerešpektoval pôvodné členenie dokumentu podľa riadkov, bolo nevyhnutné dodatočne spojiť všetky rozdelené slová. Išlo o tie slová, ktoré sa v pôvodnom dokumente nachádzali na konci riadkov. Takto pripravené dokumenty boli napokon použité ako podklad na vydanie bilingválnej pramennej edície,[36] ktorá vyšla tlačou koncom roku 2024.[37]

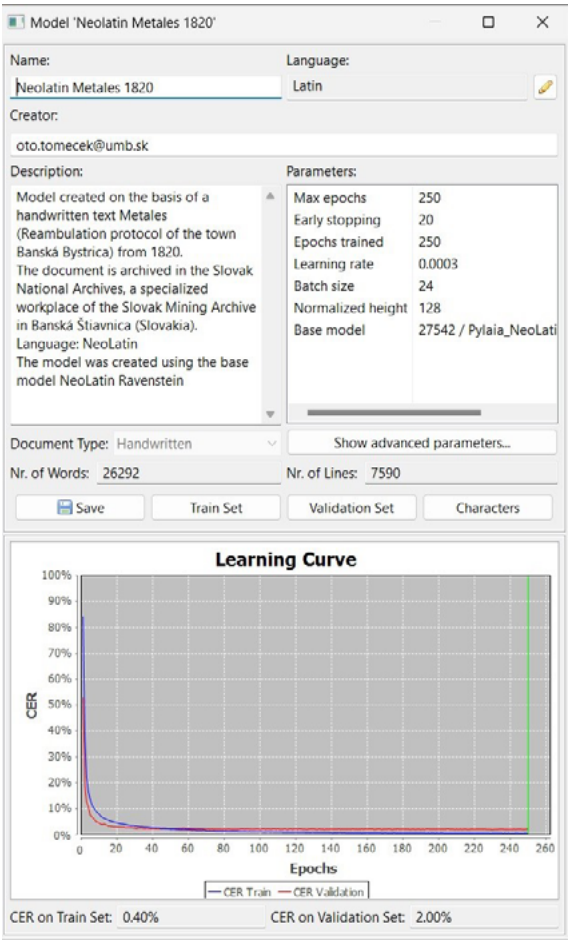
Overenie funkčnosti modelov pri transkripcii reambulačného protokolu B

Celý proces vytvárania vlastného modelu automatickej transkripcie sa ukázal ako časovo veľmi náročný. Jeho súčasťou bola realizácia manuálneho prepisu vybranej časti dokumentu, ako aj ďalšie rozširovanie tejto vzorky až na požadovaných 10 000 –15 000 slov. Práve v súvislosti s časovou náročnosťou tohto manuálneho prepisu pri tvorbe vlastného modelu sa ponúkla možnosť overiť existujúci model, vytrénovaný pre potreby transkripcie RPA, aj pre iný rukopis reambulačného protokolu B (ďalej RPB). Tento druhý rukopis protokolu totiž predstavoval ideálnu vzorku rukopisného textu, nakoľko z obsahového hľadiska je prakticky totožný s RPA.

[36] Preklad dokumentu realizoval kolega Imrich Nagy, za čo mu touto cestou rovnako vyjadrujem osobitné poďakovanie.

[37] TOMEČEK, Oto – NAGY, Imrich: *Reambulačný protokol mesta Banská Bystrica z roku 1820 : Edícia prameňa pomocou nástroja na automatickú transkripciu historických dokumentov*. Banská Bystrica 2024.

Ešte pred tým, ako som pristúpil k realizovaniu automatickej transkripcie rukopisu RPB, som vytvoril definitívny model automatickej transkripcie rukopisu RPA. Jeho vytvorenie umožnil fakt, že pre potreby prípravy pramennej edície musel byť celý prepis dokumentu znovu podrobne prekontrolovaný a opravený. Z tohto dokumentu, v celom jeho rozsahu (29 127 slov v 8 407 riadkoch), bol potom vytrénovaný *Model Neolatin Metales 1820*.[38] Pri trénovaní bol dokument v celom rozsahu rozdelený na tréningový a validačný set automaticky, približne v pomere 10:1. Tréningový set tak obsahoval 26 292 slov a 7 590 riadkov, z hľadiska výsledkov trénovania dôležitejší validačný set obsahoval 2 835 slov zapísaných v 817 riadkoch. Na týchto vzorkách bol vytrénovaný nový model, ktorý dosahoval chybovosť znakov (CER) v tréningovom sete na úrovni 0,4 % a vo validačnom sete na úrovni 2 %. Vytrénovanie modelu s takouto chybovosťou pri rukopisnom texte bolo možné považovať za excelentný výsledok a dobrý predpoklad na jeho použitie pre prepis RPB.



Obr. 6: Parametre modelu *Neolatin Metales 1820,* autor O. Tomeček

[38] V systéme *Transkribus* je tento model evidovaný pod číslom ID 59197.

Digitalizáty dokumentu RPB boli vyhotovené v študovni Štátneho archívu v Banskej Bystrici pomocou skenovacieho stanu a smartphonu – Huawei P40 Pro, ktorý nemal stiahnutú aplikáciu *DocScan app*. Týmto spôsobom boli vyhotovené snímky dokumentu s rozlíšením 96 dpi a rozmermi 4 096 x 3 072 pixelov, teda rovnakými parametrami ako v prípade väčšiny snímok hlavného súboru RPA.

Vzhľadom na časovú náročnosť kontroly a následnej opravy automatickou transkripciou prepísaného textu som sa rozhodol neoverovať model prepisom celého dokumentu, ale len jeho vybranej časti. Ako vzorku tohto experimentu som si zvolil rozsahom menšiu šiestu kapitolu RPB obsahujúcu trinásť strán textu vyhotovených na siedmich digitálnych snímkach, ktoré *Transkribus* považuje za samostatné strany (v skutočnosti dvojstrany).

Po automatickom prepise tejto vybranej vzorky RPB prostredníctvom definitívneho modelu (*Model Neolatin Metales 1820*) prebehla dodatočná kontrola a oprava chýb. Po ukončení tohto procesu bolo možné porovnať a vyhodnotiť chybovosť medzi automatickou transkripciou prepísanými stranami s ich manuálne opravenými verziami. *Transkribus* v tomto prípade vyhodnotil chybovosť na úrovni znakov aj slov po jednotlivých zdigitalizovaných stranách, v skutočnosti originálnych dvojstranách dokumentu. Chybovosť znakov (CER) na jednotlivých stranách sa pohybovala v rozmedzí od 5,68 % do 8,87 %, pričom priemerná chybovosť dosahovala 6,72 %. Chybovosť slov (WER) na jednotlivých stranách sa pohybovala v rozmedzí od 22,49 % do 37,71 %, pričom priemerná chybovosť dosahovala 28,83 %. Uvedené výsledky neboli zlé, ale ani ideálne. Práca s takto prepísaným textom by si vyžadovala ďalšiu detailnú kontrolu a opravu jednotlivých chýb. Na tomto mieste možno len upozorniť, že chybovosť na úrovni znakov (CER) by mala byť ideálne nižšia ako 10 % a na úrovni slov (WER) nižšia ako 30 %. Priemerná chybovosť prepisu strán vybranej vzorky RPB tieto hodnoty neprekračovala, takže predpoklady zrozumiteľného prepisu sa podarilo naplniť.

Druhou možnosťou overenia prepisu RPB bolo využitie supermodelu, ktorý vznikol vytrénovaním na viacerých rukopisoch a na podstatne väčšej vzorke slov. Pre vytvorenie tohto agregovaného modelu boli použité prepisy vyhotovené pomocou finálnych (najúspešnejších) modelov viacerých riešiteľov projektu Skriptor.[39] Týmto spôsobom vznikol úplne nový agregovaný model označený ako *Slovak Supermodel M1* (SSM1).[40] Do trénovania tohto modelu boli zahrnuté rukopisné písomnosti z obdobia od 16. do 20. storočia napísané v rôznych jazykoch (slovenskom, latinskom, maďarskom, českom a slovakizovanou češtinou), rôznymi typmi písma.[41] Model bol vytrénovaný na vzorke 333 777 slov, ktoré boli zapísané v 56 713 riadkoch. Pre jeho vytvorenie bolo použitých 1 359 strán upravených v kvalite *Ground Truth*. Chybovosť takto vytvoreného supermodelu na úrovni znakov (CER) dosahuje hodnotu 4,90 % v tréningovom sete a 5,30 % v smerodajnejšom validačnom sete.

Rovnakú vzorku textu RPB, ako v predošlom prípade, som teda nechal prepísať aj pomocou tohto supermodelu. Aj v tomto prípade bolo možné vyhodnotiť chybovosť

[39] Okrem agregovaného modelu určeného na prepis rukopisných textov vytvorili riešitelia projektu Skriptor aj agregovaný supermodel určený na prepis tlačených dokumentov. Tento model označený ako *Slovak Supermodel print & typewriter* (SSPT1), evidovaný pod číslom ID 78289, bol vytrénovaný na vzorke 200 697 slov s chybovosťou 1,45 % na úrovni znakov (CER) vo validačnom sete.

[40] V systéme *Transkribus* je tento model evidovaný pod číslom ID 63569.

[41] KATUŠČÁK, Dušan a kol.: Slovak Supermodel M1 (SSM1) : Matej Bel University SKRIPTOR Project Datasets (First version, 20240503) [Data set]. Zenodo. Online, cit. 20. 1. 2025, dostupné na https://doi.org/10.5281/zenodo.11109087.

automatickej transkripcie na úrovni znakov, ako aj slov. Chybovosť znakov (CER) na jednotlivých stranách sa pohybovala v rozmedzí od 4,05 % do 5,68 %, pričom priemerná chybovosť dosahovala 4,79 %. Chybovosť slov (WER) na jednotlivých stranách sa pohybovala v rozmedzí od 19,31 % do 26,82 %, pričom priemerná chybovosť dosahovala 22,57 %. Uvedené hodnoty preukázali, že prepis prostredníctvom supermodelu vykazoval v porovnaní s prepisom pomocou definitívneho modelu (*Model Neolatin Metales 1820*), vyhotoveného pre potreby prepisu RPA, vo všetkých parametroch lepšie výsledky. Priemerná chybovosť na úrovni znakov sa znížila o 1,93 %, na úrovni slov dokonca až o 6,26 %.

Strany	**Model Neolatin Metales 1820**		**Slovak Supermodel M1 (SSM1)**	
	CER	WER	CER	WER
1	6,67 %	31,73 %	5,10 %	26,44 %
2	6,77 %	26,79 %	4,39 %	19,31 %
3	6,25 %	22,49 %	5,03 %	20,41 %
4	6,94 %	30,79 %	4,69 %	23,18 %
5	5,86 %	25,44 %	4,05 %	19,53 %
6	5,68 %	26,86 %	4,58 %	22,33 %
7	8,87 %	37,71 %	5,68 %	26,82 %
Priemer	6,72 %	28,83 %	4,79 %	22,57 %

Tab. 2: Chybovosť prepisu skúšobnej vzorky RPB pomocou definitívneho modelu Neolatin Metales 1820 a slovenského supermodelu (SSM1)

Pôvodná verzia platformy *Transkribus Expert Client* umožňuje, vzájomným porovnaním vybraných verzií automaticky prepísaných a opravených strán, presne konkretizovať a graficky zvýrazniť chybne prepísané miesta textu. Tie pasáže, ktoré boli transkribované správne, ponecháva *Transkribus* bez akéhokoľvek zvýraznenia. Chybne prepísané pasáže sú označené červeným podsvietením, v ktorom je text dokumentu prečiarknutý. Hneď za ním nasleduje ten istý text v správnom znení, ktorý je naopak zvýraznený zeleným podsvietením. Takýmto spôsobom môžeme konkretizovať aj chyby vzniknuté pri automatickom prepise RPB pomocou supermodelu SSM1.

Ako zanedbateľné chyby, ktoré zásadnejšie neovplyvňujú zrozumiteľnosť prepísaného textu, možno zaradiť časté zamieňanie malých a veľkých písmen, nesprávnu identifikáciu interpunkčných znamienok, vynechanie diakritiky pri zápisoch slovenských toponým, chybnú identifikáciu medzier medzi slovami, ale napríklad aj zámenu písmen i – j, ktorá zásadnejšie nemení význam slov. Naopak, ako závažné chyby možno vyhodnotiť nepresnosti pri prepise arabských číslic, častokrát nesprávnu identifikáciu ligatúry æ, úplné vynechanie písmen alebo chyby vzniknuté takými zámenami písmen, ktoré menia štruktúru a význam slov. Pri prepise RPB prostredníctvom slovenského supermodelu boli najčastejšie zaznamenané zámeny nasledovných písmen: e – a, e – c, e – i, f – t, g – q, i – c, j – s, l – t, m – n, n – r, o – a, o – e, p – q, s – t, t – r. Okrem nich sa však zriedkavejšie objavili aj zámeny medzi inými písmenami. Aj napriek týmto uvedeným závažnejším chybám však možno konštatovať, že prepísaný text je použiteľný pre základné zorientovanie sa v obsahu dokumentu alebo fulltextové vyhľadávanie vybraných použitých slov.

Obr. 7: Ukážka vyhodnotenia chybovosti automatického prepisu RPB prostredníctvom supermodelu SSM1, autor O. Tomeček

Záver

Vyššie uvádzaná exaktná kvantifikácia výsledkov experimentov umožňuje sumarizovať a relatívne objektívne vyhodnotiť všetky podstatné zistenia v záverečnej úvahe. Použitie platformy *Transkribus* na prepis písomného historického dokumentu má viacero nespochybniteľných benefitov, ale aj viacero úskalí. Medzi výhody tohto konkrétneho nástroja patrí možnosť vytvárať vlastné modely a tieto ďalej vylepšovať ich trénovaním. Tento postup je vhodné uplatniť hlavne pri rozsiahlejších textoch s veľkým počtom slov napísaných jednou písarskou rukou. Ďalšou výhodou je možnosť použitia takzvaných *Base* modelov, ktorých použitie je však efektívne hlavne v kombinácii s vlastným modelom. Za obrovský benefit možno považovať vytváranie a využívanie veľkých agregovaných modelov vytrénovaných na veľmi veľkom počte slov rozličných rukopisov. Ich použitie sa ukazuje ako mimoriadne efektívne hlavne pri kratších textoch, resp. textoch napísaných viacerými písarskymi rukami. Negatívne môže vyznievať časová náročnosť niektorých nevyhnutných krokov predchádzajúcich samotnej automatickej transkripcii. V tomto prípade ide predovšetkým o proces segmentácie dokumentu, teda procesu, s ktorým sa klasický prekladateľ pri bežnom mechanickom preklade prakticky nestretáva. Práve proces segmentácie je teda zvyčajne tým rozdielovým elementom, ktorý rozhoduje pri dileme o prípadnej časovej úspore využitia nástroja automatickej transkripcie. Odhliadnuc od individuálnych zručností a schopností jednotlivca, aj v tomto prípade veľa závisí

práve od charakteru dokumentu určeného na automatickú transkripciu. Pri jednoduchých a vnútorne nečlenených textoch je možné uplatniť automatickú segmentáciu, ktorá pri následnej kontrole zvyčajne vyžaduje len menšie korekcie. Naopak, pomerne zdĺhavým sa tento proces stáva pri dokumentoch spracovaných v tabuľkovej forme, prípadne obsahujúcich množstvo vsuviek a marginálií, kde je nevyhnutné realizovať mechanickú segmentáciu.

Viaceré vyššie spomenuté závery možno zovšeobecniť ako platné pri použití akéhokoľvek nástroja automatickej transkripcie. Pri rozsiahlejších textoch napísaných jednou pisárskou rukou sa ukazuje ako účelné vytváranie vlastného modelu ušitého priamo na mieru konkrétnemu dokumentu. V prípade kratších textov, resp. textov napísaných viacerými pisárskymi rukami, sa ako účelnejšie javí naopak používanie dostupných supermodelov vytrénovaných na veľkom počte rukopisov. Ukazuje sa, že budúcnosť automatickej transkripcie sa nachádza práve vo vytváraní supermodelov vytrénovaných na viacerých rukopisoch a veľkom počte slov. Ideálnym riešením do budúcnosti sa javí vytrénovanie takýchto supermodelov prispôsobených konkrétnemu typu písma, prípadne konkrétnemu obdobiu. Osobitne by takto mohol vzniknúť napríklad supermodel pre potreby prepisu rôznych druhov tlačeného písma, strojopisu, ale primárne predovšetkým na rôzne typy rukopisov, napríklad stredoveké kancelárske písmo, humanistický kurent, humanistická kurzíva a podobne. Vytrénovaním takýchto supermodelov sa stane aj používanie akéhokoľvek nástroja automatickej transkripcie podstatne jednoduchšie, účelnejšie a časovo menej náročné.

Zostáva len veriť, že tento príspevok, založený na vlastných skúsenostiach s prácou s konkrétnym nástrojom určeným na automatickú transkripciu rukopisných textov, vniesol viac svetla do tejto aktuálnej a v historickej obci mimoriadne diskutovanej problematiky. Skoncipovanie predloženého textu možno vnímať ako snahu vytvoriť akúsi pomôcku pri riešení dilemy o možnom využití *Transkribu*, ako nástroja automatickej transkripcie, vo vlastnej práci s historickým rukopisným textom. Na základe nej môže každý bádateľ prijať vlastné rozhodnutie či vôbec, resp. za akých okolností, je vhodné tento pracovný nástroj použiť.

Bibliografia

Archívne pramene

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

Internetové a elektronické zdroje

<https://readcoop.eu/scantent/>
<https://readcoop.eu/Transkribus/>
<https://readcoop.eu/Transkribus/resources/how-to-guides/>
<https://www.youtube.com/watch?v=KLtLg5Dui80 & list=PL7UbQtd4qlhI3VJck98N4kDAeuZfy--H0>
<https://www.youtube.com/watch?v=LdBYSDMSeC8>
<https://www.youtube.com/watch?v=3mv7_pYDK5E>

KATUŠČÁK, Dušan a kol.: Slovak Supermodel M1 (SSM1) : Matej Bel University SKRIPTOR Project Datasets (First version, 20240503) [Data set]. Zenodo. Dostupné na https://doi.org/10.5281/zenodo.11109087.

Literatúra

BURDICKOVÁ, Anne a kol.: *Digital Humanities*. Praha 2019.
FOLTÝN, Tomáš: Digital Humanities : stručné shrnutí stávajícího stavu problematiky v ČR. *ITlib : Informačné Technológie a Knižnice* 21, 2017, č. 3, s. 19–22.
GOGORA, Andrej: Ako a prečo pomenovať to, čo robíme? : Problém slovenského prekladu termínu „digital humanities". *Slovenská literatúra* 67, 2020, č. 6, s. 598–613.
JANNIDIS, Fotis – KOHLE, Hubertus – REHBEIN, Malte: *Digital Humanities : Eine Einführung*. Stuttgart 2017.
KATRENIAK, Martin: *Automatická transkripcia rukopisných historických textov na príklade vybraných kanonických vizitácií*. Diplomová práca. Banská Bystrica 2022.
KATRENIAK, Martin – KUNEC, Patrik: Automatická transkripcia historických prameňov obsahujúcich viac rukopisov na príklade kanonických vizitácií. In: MALINIAK, Pavol – NAGY, Imrich (eds.): *Digital humanities : Nástroje sprístupňovania historického dedičstva*. Banská Bystrica 2022, s. 48–50.
KATUŠČÁK, Dušan: Digital humanities a automatická transkripcia rukopisných textov. *ITlib : Informačné Technológie a Knižnice* 24, 2020, č. 1, s. 6–16.
KATUŠČÁK, Dušan – NAGY, Imrich (eds.): *Automatická transkripcia historických dokumentov : metodická príručka na prácu s platformou Transkribus*. Banská Bystrica 2023.
KATUŠČÁK, Dušan – NAGY, Imrich (eds.): *Automatická transkripcia slovacikálnych historických dokumentov*. Banská Bystrica 2023.
MALINIAK, Pavol – NAGY, Imrich (eds.): *Digital humanities : Nástroje sprístupňovania historického dedičstva*. Banská Bystrica 2022.
MILLIGAN, Ian: *The Transformation of Historical Research in the Digital Age*. Cambridge – New York – Melbourne – New Delhi – Singapore 2022.
MUEHLBERGER, Guenter a kol.: Transforming Scholarship in the Archives through Handwritten Text Recognition : Transkribus as a Case Study. *Journal of Documentation* 75, 2019, no. 5, s. 954–976.
NAGY, Imrich: Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. *Slovenská archivistika* 51, 2021, č. 2, s. 53–67.
NAGY, Imrich – KATUŠČÁK, Dušan (eds.): *Transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus : metodická príručka pre účastníkov workshopu*. Banská Bystrica 2024.
TOMEČEK, Oto: Automatická transkripcia reambulačného protokolu Banskej Bystrice z roku 1820. In: KATUŠČÁK, Dušan – NAGY, Imrich (eds.): *Automatická transkripcia slovacikálnych historických dokumentov.* Banská Bystrica 2023, s. 102–123.
TOMEČEK, Oto – NAGY, Imrich: *Reambulačný protokol mesta Banská Bystrica z roku 1820 : Edícia prameňa pomocou nástroja na automatickú transkripciu historických dokumentov*. Banská Bystrica 2024.

Summary

On the Possibilities of Automatic Transcription of Historical Handwritten Documents Using the Transkribus Platform

To facilitate the historian's work with written sources, the Transkribus platform, which is primarily designed for the automatic transcription of historical handwritten documents, can now be used. The added value of this artificial intelligence tool is the ability to train custom models designed to transcribe a specific document. Creating and training custom models is quite time-consuming. The author of the study tried all the steps of the automatic transcription in the environment of the Transkribus platform while working with a specific document, the reambulatory protocol of the town of Banská Bystrica from 1820. While preparing for the implementation of the automatic transcription, he also trained his own model using the basic Neolatin Ravenstein model. The final automatic transcription was implemented using the custom model, which achieved a character error rate of 2.60 % in the validation set.

After the automatic transcription of the entire document and the subsequent correction of all errors, the final Neolatin Metales 1820 model was generated with an error rate of 2 % at the character level in the validation set. On the basis of this successful model with an extremely low error rate, the author decided to test this model by transcribing another manuscript of the same reambulatory protocol. The result of this experiment was less satisfactory. The selected sample of the other ream-

bulatory protocol was transcribed by Transkribus using the above model with an average error rate of 6.72 % at the character level and 28.83 % at the word level. The next experiment was to transcribe an identical sample of text using the Slovak Supermodel M1 (SSM1), trained on a much larger sample of handwritten documents. In this case, the transcription results were better than those of the previous one, as it managed to achieve an average error rate of 4.79 % at the character level and 22.57 % at the word level.

The results of the above experiments can be interpreted in such a way that the creation of custom models is particularly useful for larger texts written by one hand. On the other hand, for shorter texts or texts written by several scribes, the use of available supermodels trained on a large number of manuscripts seems to be more efficient. It turns out that the future of automatic transcription lies precisely in the creation of such supermodels, trained on multiple handwritings and a large number of words. The ideal solution for the future seems to be the training of such supermodels, adapted to the particular type of writing used in a particular period. Only when a number of such supermodels have been trained will the transcription of handwritten documents by means of automatic transcription become a truly effective and practical aid to the historian's work.