Univerzálne modely automatickej transkripcie historických tlačí: Experimenty s platformou Transkribus na tlačiach *Orbis Pictus* (1798) a *Adparatus ad Historiam Hungariae* (1735) [1]

LUCIA NIŽNÍKOVÁ – MICHAELA MIKUŠKOVÁ

Univerzitná knižnica Univerzity Mateja Bela v Banskej Bystrici

Universal Transcription Models for Historical Prints: Experiments with Transkribus on *Orbis Pictus* (1798) and *Adparatus ad Historiam Hungariae* (1735) [2]

Abstract: This paper contributes to digital humanities research and demonstrates the use of Artificial Intelligence in the processing of written cultural heritage. It focuses on the practical application of HTR+ (Handwritten Text Recognition) tools and the Transkribus platform in the context of automatic recognition and transcription of historical prints, and examines their usability in creating universal transcription models. The aim of the paper is not to analyze or compare various tools for the transcription of handwritten texts and historical prints, but to make the evaluation of several years work with Transkribus software based on the results achieved by the authors as part of the applied research project *SKRIPTOR* (2020 – 2024). This research was carried out with the *Transkribus Expert Client* and the *Transkribus App* software using HTR+ and PyLaia technologies. The procedure included digitizing the Orbis Pictus (1798) textbook using the *ScanTent* and *DocScan* tools, manual text segmentation, preparation of the Ground Truth sample, and training automatic transcription models for four font types (antiqua, cursive, Fraktur, Schwabacher) and four languages (Latin, Hungarian, German, Czech). The best trained models were verified on historical prints of *Orbis Pictus* (1820) and *Adparatus ad Historiam Hungariae* (1735). Their universality was evaluated using the CER and WER indicators. The research also pointed to limiting factors in the quality of the digitized material, the layout of the text, and the appearance of specific graphemes. Testing the Transkribus Print M1 supermodel demonstrated its high accuracy and usability for historical multilingual prints.
Keywords: automatic transcription, historical prints, Transkribus, digital humanities.

DOI: https://doi.org/10.24040/ahn.2026.29.01.93-122

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

[2] This paper was supported by the Slovak Research and Development Agency under contract no. APVV-19-0456 *Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts (SKRIPTOR)*.

Aktuálne v spoločnosti prebieha debata o tom, ako umelú inteligenciu uchopiť a čo najefektívnejšie aplikovať do praxe. Možnosťami jej využívania vo výskume sa zaoberajú vedci naprieč celým spektrom vedných odborov. Prvé náznaky prepájania humanitných vied a výpočtovej techniky možno spätne vystopovať až k priekopníckej práci z konca 40. rokov 20. storočia. Od konca 80. rokov do začiatku 21. storočia sa rozvinula prvá vlna digitálnych humanitných vied, ktorá sa zaoberala spôsobmi štruktúrovania humanitných dát s cieľom efektívne komunikovať s výpočtovou technikou [3]. *Digital humanities* dnes prinášajú nové formy výskumu, výučby a publikovania vrátane kolaboratívnych a transdisciplinárnych prístupov.

Umelá inteligencia dokáže napodobňovať ľudské myslenie, učiť sa na základe veľkého množstva dát a čo je najdôležitejšie, zlepšovať svoje výkony v závislosti od získaných skúseností. Na tejto schopnosti sú založené európske projekty *tranSciptorium* [4], riešený v rokoch 2013 – 2015 a financovaný zo 7. rámcového programu, a projekt *READ – Recognition and Enrichment of Archival Documents* [5], realizovaný v rokoch 2016 – 2019 a financovaný zo zdrojov programu Horizont 2020 [6]. Ich výsledkom je inovatívny nástroj *Transkribus*, ktorý umožňuje automatické rozpoznávanie textu a transkripciu historických dokumentov. Predpokladom jeho využitia v praxi sú za pomoci umelej inteligencie vytrénované modely na prepis rôznych rukopisných štýlov a typov písma (tlačeného a strojopisného) vrátane špecifických grafém a rukopisných skratiek, a to bez ohľadu na jazyk dokumentu.



Obrázok 1. Ponuka platformy Transkribus. Zdroj: https://readcoop.eu/transkribus

Metódy rozpoznávania rukopisných textov (HTR – Handwritten Text

[3] BURDICK, Anne – DRUCKER, Johanna – LUNENFELD, Peter et al. *Digital Humanities* [online]. The MIT Press, 2012, pp. 8 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.7551/mitpress/9248.001.0001

[4] TranScriptorium, https://cordis.europa.eu/project/id/600707

[5] Recognition and Enrichment of Archival Documents, https://cordis.europa.eu/project/id/674943

[6] MÜHLBERGER, Günter. *H2020 Project READ (Recognition and Enrichment of Archival Documents) – 2016 – 2019* [online]. 2021 [cit. 2026-03-26]. Dostupné na: https://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019

Recognition) využívajú umelé neurónové siete (ANN – Artificial Neural Networks) ako jeden z algoritmov strojového učenia. Ich využitie v praxi funguje na základe vytvorenia základného modelu, ktorý sa cvičí na vybranej vzorke dát. Na základe postupného zväčšovania objemu dát sa trénovaný model zdokonaľuje a zvyšuje sa jeho presnosť. Transkribus ponúka stovky vytrénovaných verejných modelov, ktoré boli vyškolené na určitý typ písma, rukopisu a typ dokumentu. Hoci bol primárne vyvíjaný ako nástroj na automatickú transkripciu rukopisných textov, veľmi dobré výsledky vykazuje aj na tlačiach a strojopisných zbierkach[7]. Po ukončení pilotného projektu READ sa na jeho udržateľnosti a na ďalšom vývoji podieľala organizácia READ-COOP SCE združujúca rôzne vedecké, výskumné, vzdelávacie a archívne inštitúcie. Do roku 2024 boli dostupné dve verzie Transkribusu:

- *Transkribus Expert Client* – pôvodná softvérová verzia, ktorá sa inštaluje priamo do osobného počítača; od roku 2024 sa nepočíta s jej ďalším vývojom, ale s postupným presunom jej funkcií do webovej verzie[8],
- *Transkribus Lite* (neskôr premenovaná na *Transkribus app*) – webová verzia, na ktorú je zameraný súčasný a budúci vývoj.

Obe platformy ponúkajú aj nástroje na editáciu, spracovanie a tagovanie transkribovaných textov. Výsledky transkripcie bolo spočiatku možné sprístupniť verejnosti na portáli *Read&Search*[9], platenej platforme Transkribusu, ktorý sprístupňoval dokumenty zo zbierky vytvorenej na platforme Transkribus Expert Client online formou. Po prechode klienta na webovú verziu na tento účel slúži prostredie *Transkribus Sites*[10]. Rozhranie bohaté na funkcie je ideálne na sprístupnenie historických dokumentov a vyhľadávanie na webe. Alternatívne je možné, po exporte v podporovaných formátoch, výstupy transkripcie a zdrojové obrázky sprístupňovať aj na iných webových sídlach alebo v digitálnych repozitároch. Pridanou hodnotou je možnosť exportovať dokument aj podľa tagov.

Na tieto projekty nadväzuje slovenský projekt SKRIPTOR z rokov

[7] SMIDA, Matej. Možnosti automatickej transkripcie v platforme Transkribus na príklade správy o vybavovaní sťažností občanov v období komunistickej diktatúry. In: *Acta historica Neosoliensia* [online]. 2023, roč. 26, č. 1, s. 125-148 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/ahn.2023.26.01.125-148

[8] V lete 2023 vývojový tím Transkribus avizoval ukončenie vývoja funkcií v prostredí expert klienta a zameranie vývoja na online verziu Transkribus Lite/Beta. Na rozvoj online platformy a presun funkcií dostupných zatiaľ len v expert klientovi boli nasmerované všetky vývojové aktivity. Napriek tomu ešte v januári 2024 mali používatelia expert klienta dostupný upgrade na novšiu verziu 1.27.0. Definitívne ukončenie platformy Transkribus Expert Client bolo napokon naplánované na jún 2024. Pre nových používateľov už nie je k dispozícii.

[9] https://readcoop.eu/readsearch/

[10] https://www.transkribus.org/sites#features

2020 – 2024. Predmetom výskumu boli historické dokumenty, ich automatická transkripcia a uplatnenie digitálnych a informačných technológií v oblasti humanitných vied. Výskumným problémom projektu bola tvorba čo najlepších modelov automatického rozpoznávania textov historických slovacikálnych dokumentov, ktoré predstavujú súčasť európskeho písomného dedičstva a prezentujú našu kultúru v kontexte európskej vedy, kultúry, politiky, hospodárstva a vzdelávania. Špecifickým výskumným problémom projektu bola digitalizácia, atraktívna prezentácia a inovatívne sprístupnenie týchto dokumentov širokej odbornej a laickej verejnosti.

Na účely aplikovaného výskumu projektu Skriptor bola použitá platforma *Transkribus Expert Client*. V prvej polovici obdobia trvania projektu boli v expert klientovi na trénovanie modelov a automatickú transkripciu dokumentov k dispozícii dve technológie:

- *HTR+* vyvinutá University of Rostock,
- *PyLaia* vyvinutá Universitat Politècnica de València.

PyLaia je zdokonalením pôvodnej HTR+ technológie a používateľom umožňuje ľubovoľné nastavenie rôznych parametrov sieťovej štruktúry pri trénovaní modelu. Podpora HTR+ bola zo strany vývojového tímu Transkribus ukončená v novembri 2022. Táto zmena sa výrazne dotkla dovtedy vytrénovaných modelov riešiteľského kolektívu.

V súčasnosti existuje už celý rad nástrojov, ktoré slúžia na automatické rozpoznávanie zdigitalizovaných, pôvodne tlačených dokumentov, medzi nimi OCR4all, eScript, Rescribe, Pero.cz, ABBYY Cloud OCR SDK, Online OCR, Kofax, Omnipage, a iné. Zámerom tohto článku nie je analýza, ani porovnávanie týchto nástrojov, ale zhodnotenie využiteľnosti softvéru Transkribus s možnosťami a funkcionalitami, ktoré aktuálne ponúka. V rámci aplikovaného výskumu projektu Skriptor sa riešiteľky zamerali na automatickú transkripciu unikátnych tlačených dokumentov v jazykoch nášho kultúrneho regiónu, medzi ktoré patrí slovenčina, čeština, latinčina či maďarčina, a na možnosti ich sprístupnenia. Na tento účel boli vybrané dve tlače: učebnica J. A. Komenského *Orbis Pictus* (1798) a dielo Mateja Bela *Adparatus ad Historiam Hungariae* (1735).

Prvé štvorjazyčné vydanie Komenského učebnice *Orbis Pictus*[11] z dielne prešporského tlačiara Šimona Petra Webera z roku 1798 je súčasťou historického knižničného fondu Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici. Na účely projektu bola učebnica zaujímavá práve tým, že ide o štvorjazyčné vydanie, pričom každý jazyk je vytlačený iným typom písma – latinčina antikvou, maďarčina kurzívou, nemčina fraktúrou a čeština švabachom. Podobný výskum, avšak na rukopisnom prameni,

[11] COMENII, Joann Amos. *Orbis pictus, in hungaricum, germanicum et slavicum translatus et hic ibive emendatus.* Posonii: Sumtibus & Typis Simonis Petri Weber, 1798.

realizovali Capurro et al.[12] Cieľom aplikovaného výskumu bolo vytrénovať použiteľný univerzálny model automatickej transkripcie pre štyri jazyky a štyri typy písma. Stoosemdesiatštyristranová publikácia s rozmermi 18,9 x 11,5 cm je vytlačená na ručne odlievanom papieri. V textovej časti sú početne zastúpené obdĺžnikové drevoryty[13]. 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.

Práca s dokumentom na platforme Transkribus pozostáva z niekoľkých fáz. Vytvorenie konta v *Transkribus app* je bezplatné. Po prihlásení získava používateľ, ktorý chce platformu využívať zdarma, každý mesiac voľných 50 kreditov. Táto forma prihlásenia však umožňuje prístup len k obmedzenej ponuke funkcionalít. Prístup k ďalším nástrojom sa v ďalších typoch predplatného (*Scholar*, *Team*, *Organisation*)[14] zvyšuje priamo úmerne s poplatkami. Kredity sa aktuálne používajú na segmentáciu dokumentu a na trénovanie modelov, ostatné práce v aplikácii sú zatiaľ bezplatné.

Zhotovenie digitalizátu a nahratie na platformu

Základným predpokladom na prácu s Transkribusom je zhotovenie digitalizátu dokumentu, ktorý bude slúžiť na trénovanie modelu a následnú transkripciu, prípadne na transkripciu s použitím už vytrénovaného vlastného alebo verejného modelu. Transkribus momentálne pracuje s obrázkami vo formátoch JPEG/JPG a PNG a so súbormi formátu PDF. Snímky dokumentu by mali byť vytvorené vo vývojármi odporúčanej kvalite priemerne 300 DPI. Najvyššiu kvalitu snímania umožňujú profesionálne skenery (400 – 600 DPI), ale takéto vysoké rozlíšenie je

[12] CAPURRO, Carlotta – PROVATOROVA, Vera – KANOULAS, Evangelos. Experimenting with training a neural network in transkribus to recognise text in a multilingual and multi-authored manuscript collection. In: *Heritage* [online]. 2023, vol. 6, no. 12, pp. 7482–7494 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.3390/heritage6120392

[13] BÔBOVÁ, Mária – KATUŠČÁK, Dušan – KURHAJCOVÁ, Alica et al. *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2022, s. 168-192 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2022.9788055720203

[14] Cenová a kreditová politika na platforme sa v posledných rokoch mení, preto si treba odsledovať aktuálne platné podmienky. Kredity sa v procese automatickej transkripcie textu priebežne odpočítavajú v závislosti od parametrov nastavených pred spustením automatickej transkripcie, aj od rozsahu a typu dokumentu, ktorý sa prepisuje (historická tlač, strojopisný text, rukopisný dokument). Viac informácií o poplatkoch pre rôzne typy predplatného na https://www.transkribus.org/plans

zbytočné, pretože nijako neprispieva k zlepšeniu rozpoznania a automatickej transkripcie textu. Transkribus vykazuje vynikajúce výsledky aj s obrázkami nafotenými vlastným zariadením, napr. fotoaparát, mobilný telefón, tablet a i., pokiaľ disponujú kvalitnou optikou. Na uľahčenie snímania dokumentu vo vyhovujúcej kvalite bola pre potreby projektu READ vyvinutá prenosná pracovná pomôcka *ScanTent* s vlastným podsvietením zabezpečujúcim rovnomerné osvietenie dokumentu (nevýhodou je obmedzený rozmer plochy určenej na snímanie do veľkosti A3) a aplikácia DocScan voľne stiahnuteľná do smartfónu, ktorá umožňuje snímanie dokumentu aj bez neustáleho stláčania spúšte fotoaparátu. Na digitalizáciu tlače Orbis Pictus (1798) sme využili pomôcku ScantTent a aplikáciu DocScan, výsledkom bolo deväťdesiatštyri snímok dvojstrán s rozlíšením 72 DPI[15].

Nevyhnutnou podmienkou pre vkladanie digitalizátov na platformu je vytvorenie vlastnej zbierky. Do nej sa následne pomocou funkcie *+Upload* nahrajú zdigitalizované snímky dokumentu v niektorom z predpísaných formátov. Ak sa snímky nahrávajú jednotlivo vo formátoch JPEG a PNG, nemali by presahovať veľkosť 10 MB. V prípade nahrávania dokumentu v PDF formáte by jeho veľkosť nemala byť väčšia ako 200 MB a maximálny počet strán v súbore by nemal prekročiť hodnotu 3000[16]. Na tomto mieste sa žiada zdôrazniť, že všetky snímky a modely, ktoré sa nachádzajú v zbierke používateľa, sú – hoci uložené na serveroch platformy – v jeho výlučnom vlastníctve. Každý používateľ má právo rozhodnúť sa, či, do akej miery a komu umožní prístup do svojej zbierky.

Segmentácia textu

Prvým krokom práce s naimportovanými digitalizátmi je ich segmentácia, ktorou sa určuje orientácia a poradie čítania textu na snímkach. Zahŕňa vymedzenie textových blokov (Text Regions) a riadkov v nich (Baselines) a má vplyv na kvalitu trénovaného modelu a následný automatický prepis dokumentu. Segmentáciu je možné vykonať automaticky

[15] Pri ďalších prácach na výskume sme použili vydanie tlače z iného roku a v dvoch rôznych rozlíšeniach: Orbis Pictus z roku 1820 z fondu Univerzitnej knižnice UMB zdigitalizovaný na profesionálnom skeneri v Centre vedecko-technických informácií SR v Bratislave v rozlíšení 600 DPI a Orbis Pictus z roku 1820 stiahnutý z portálu Digitálna knižnica a digitálny archív Slovenskej národnej knižnice (DIKDA) v rozlíšení 96 DPI.

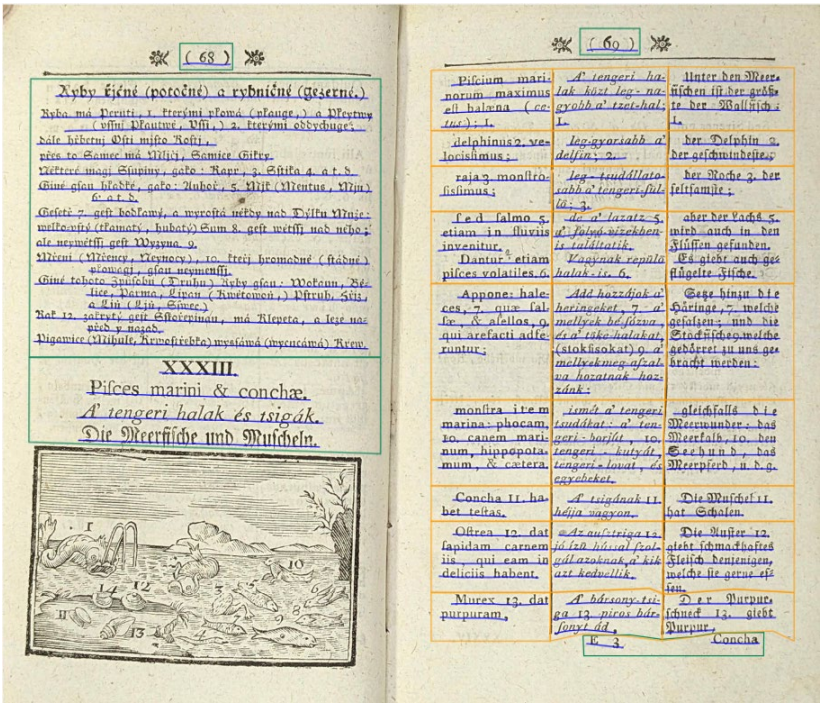
[16] BÔBOVÁ, Mária – KATUŠČÁK, Dušan – KURHAJCOVÁ, Alica et al. *Automatická transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus: Metodická príručka pre účastníkov workshopu* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2024, s. 41-45 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2024.9788055721439

alebo manuálne. Vzhľadom na štruktúru dokumentu Orbis Pictus (kombinácia klasického knižného textu a stĺpcového členenia) sme pristúpili k manuálnemu označeniu textových rámcov zohľadňujúcich členenie textu, jeho jazykové mutácie a typy písma. V zadefinovaných textových rámcoch sme spustili automatickú detekciu riadkov. Ako vo svojom článku upozorňuje Ost[17], automatická segmentácia nie je bezchybný proces. Najmä pri štruktúrovaných dokumentoch zvyčajne treba urobiť kontrolu a vykonať väčšie, alebo menšie korekcie v delení riadkov a správnom čítaní ich poradia[18]. Technologické zmeny na platforme Transkribus (november 2022) priniesli nové funkcionality a zefektívnenie niektorých jestvujúcich. Tie sa prejavili najmä na úrovni segmentácie dokumentu, keď pôvodnú metódu analýzy textu CITlab Advanced LA nahradila metóda Transkribus LA. Tá ponúka možnosť nastavenia viacerých parametrov delenia riadkov a ich automatického rozdeľovania, ak sa text nachádza v tesnej blízkosti hranice textového rámca, čo bol aj prípad dokumentu Orbis Pictus.

Na trénovanie modelu a automatický prepis dokumentu vplýva správne zadefinovanie základných čiar (Baselines). Tie predstavujú najdôležitejší referenčný bod na rozpoznávanie textu. Opisujú líniu (polyčiaru), ktorá sa tiahne pozdĺž celej spodnej časti riadka. Pri automatickej segmentácii riadkov je potrebné hranice začiatku a konca základnej čiary manuálne korigovať. Podľa definovanej základnej čiary sa vytvorí oblasť riadka, ktorá predstavuje polygón zahŕňajúci text dokumentu na príslušnom riadku. Korekcia základných čiar v rozsiahlejších dokumentoch môže byť časovo náročná. Menší rozsah tlače Orbis Pictus nám však umožnil detailné porovnanie správnosti prepisu chybne určených základných čiar. Chyby v určení základnej čiary boli rôzneho rozsahu (nezahŕňali celý znak, jeho časť, diakritické znamienka, rozdeľovníky a i.), pričom sa vyskytovali na začiatku i na konci riadka.

[17] OST, Katharina. Möglichkeit und Grenzen einer groß angelegten Volltexterschließung von Inkunabeln mit Transkribus. In: *Zeitschrift Für Bibliothekswesen Und Bibliographie* [online]. 2024, vol. 71, no. 3, pp. 169-181 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.3196/186429502471338

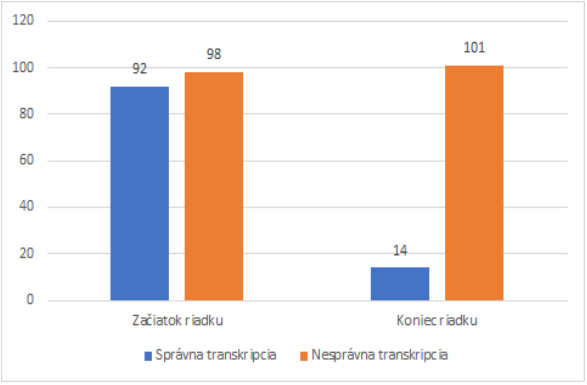
[18] BÔBOVÁ, Mária – KATUŠČÁK, Dušan – KURHAJCOVÁ, Alica et al. *Automatická transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus: Metodická príručka pre účastníkov workshopu* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2024, s. 72-76 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2024.9788055721439



Obrázok 2. Ukážka segmentácie tlače Orbis Pictus. Zdroj: Transkribus

Z nižšie uvedeného grafu vidieť, aký vplyv má nesprávne definovaná línia základnej čiary na prepis. Softvér pri transkripcii pomocou použitia vybraného modelu dokáže v porovnaní so stanovenou základnou čiarou zachytiť širšiu oblasť riadka. V takom prípade môže mať oblasť riadka pozitívny vplyv na správny prepis grafém na hranici začiatku a konca základnej čiary.

Graf 1. Chybovosť transkripcie pri nesprávne určenej línii základnej čiary na začiatku a konci riadka



Potvrdilo sa tiež, že nesprávne nadefinovaný polygón označujúci text riadka nemá pri správne zadefinovanej základnej čiare vplyv na prepis textu. Chybovosť prepisu pri nesprávne určenej línii základnej čiary bola vyššia na konci riadka ako na jeho začiatku.

Príprava vzorky *Ground Truth*

Po segmentácii textu sa na automatickú transkripciu dokumentu použije starostlivo vybraný model iného používateľa[19] alebo sa vytrénuje nový model špecifický pre dokument, ktorý je predmetom skúmania. Pred trénovaním nového modelu automatickej transkripcie je potrebné pripraviť vzorku dát, ktorá sa použije na trénovanie. V strojovom učení sa na označenie takýchto dát používa termín *Ground Truth (GT)*[20]. Strany možno prepísať manuálne alebo použiť vhodný model iného používateľa. Následne sa tento prepis dôsledne skontroluje, skoriguje a strany sa označia ako *Ground Truth*. Druhý spôsob významne šetrí ľudskú prácu

[19] Použiť sa dá voľne dostupný model, ktorý najlepšie zodpovedá typu a jazyku dokumentu, použitému fontu, prípadne obdobiu jeho vzniku. Aktuálne sú na platforme dostupné aj supermodely *The Text Titan I* a *II*, ktoré však platforma odporúča používať na automatickú transkripciu rukopisných dokumentov.

[20] *Ground Truth* = overená, skutočná informácia použitá na trénovanie a testovanie modelov umelej inteligencie a na porovnanie s výsledkami týchto modelov. Je to základný referenčný bod, ktorý model porovnáva s vlastnými predikciami, aby overil svoju presnosť a naučil sa identifikovať vzory.

i čas a bol použitý pri viacerých výskumoch [21]. Odporúčaný počet transkribovaných slov pre tlačené dokumenty [22] je minimálne 5 000.

V prípade historických textov metodika na prácu so softvérom Transkribus odporúča ísť cestou diplomatického prepisu, ktorý presne odráža to, čo sa nachádza v pôvodnom dokumente [23]. V závislosti od vôle zachovať čo najvyššiu mieru autenticity dokumentu je preto dôležité vopred sa rozhodnúť, či k prepisu dokumentu budete pristupovať metódou:

- *transkripcie* – písomného vyjadrenia (vyslovovaných alebo cudzím grafickým systémom napísaných) slov a textov z hľadiska ich výslovnosti prostriedkami určitého grafického systému [24], teda počíta s nahradením starších a nepoužívaných znakov modernými, alebo
- *transliterácie* – prevodu 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 [25], čiže prepisu znakov v ich pôvodnej podobe.

Štvorjazyčnosť dokumentu, s ktorým sme pracovali, a použité znakové sady (fonty) poskytujú zaujímavý materiál na ďalšie analýzy (grafologický a jazykový výskum), preto sme uplatnili metódu transliterácie (aj s chybami, ktoré vznikli pri tlači dokumentu). Na prepis sa v Transkribuse používajú bežné sady UNICODE, pričom softvér obsahuje aj funkciu na import nových znakov. Pri manuálnom prepise špecifických znakov, ktoré softvér neobsahoval, sme spočiatku používali náhradné znaky, ktoré boli významovo a foneticky podobné [26]. Na vyladenie transliterácie sme nakoniec využili nástroj na tvorbu kombinácií UNICODE znakov. Každý jazyk a font obsahoval grafémy, z ktorých mnohé sa v dnešnej

[21] CAPURRO, Carlotta – PROVATOROVA, Vera – KANOULAS, Evangelos. Experimenting with training a neural network in transkribus to recognise text in a multilingual and multi-authored manuscript collection. In: *Heritage* [online]. 2023, vol. 6, no. 12, pp. 7482–7494 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.3390/heritage6120392 a MARSILI, Giulia – HASSAM, Stephan N. From archives to museum and back: Transcribing, digitizing, and enriching cultural heritage and manuscript legacy data of the Villa del Casale of Piazza Armerina. In: *Digital Applications in Archaeology and Cultural Heritage* [online]. 2025, vol. 38, e00441 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.1016/j.daach.2025.e00441

[22] Viac o príprave dát na trénovanie modelu https://help.transkribus.org/data-preparation

[23] Viac ku konvenciám manuálneho prepisu https://help.transkribus.org/data-preparation

[24] MISTRÍK, Jozef. Prepis z iného písma. In: *Encyklopédia jazykovedy*. Bratislava: Obzor, 1993, s. 343.

[25] MISTRÍK, Jozef. Prepis z iného písma, s. 343.

[26] BÔBOVÁ, Mária – KATUŠČÁK, Dušan – KURHAJCOVÁ, Alica et al. *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2022, s. 168-192 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2022.9788055720203

dobe už nepoužívajú. Ich najvyšší výskyt bol v českom jazyku:

- latinčina (antikva) → æ, œ, ë
- maďarčina (kurzíva) → ő, ű
- nemčina (fraktúra) → ß, ʒ, å, ů, ů
- čeština (švabach) → ċ, ė, ġ, ṅ, ṙ, ṡ, ṫ, ż, ě, ů

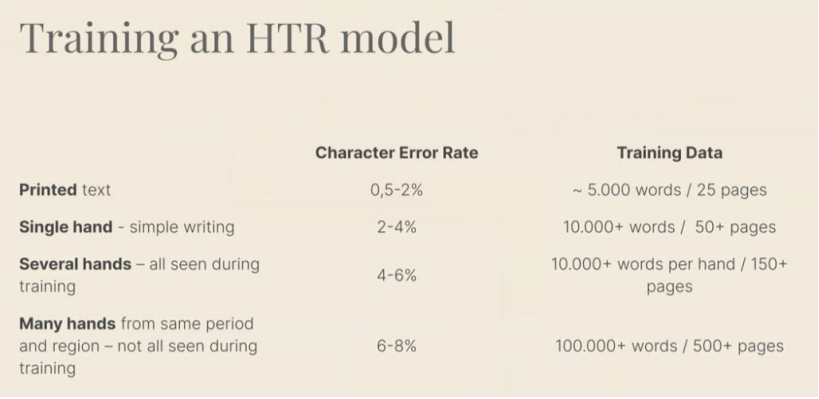
Spoluhlásky v češtine sa vyskytovali aj vo forme kapitálok. Všetky fonty obsahovali grafému dlhé s (ſ), ktorá nahrádza okrúhle s na začiatku a uprostred slova.

Na to, aby model dosiahol čo najlepšie výsledky, je nevyhnutné, aby vzorka GT obsahovala dostatočné zastúpenie všetkých znakov. Potvrdilo sa to aj v prípade tlačených dokumentov. Pri trénovaní modelu na tlači *Orbis Pictus* (1798) boli v cvičnom súbore zastúpené znaky ř, ě, č, ë, ġ, Ġ, œ. Správnosť prepisu týchto znakov nebolo možné dostatočne overiť na overovacom súbore trénovaného modelu, pretože sa v ňom nemuseli vyskytovať. Preto sme pristúpili k analýze na celom transkribovanom texte. Ich prepis bol prevažne nesprávny, najčastejšie išlo o zámenu œ → æ, ë → e/ċ, ġ → g, ř → r/ṙ/f/k, č → c/ċ/e.

Nastavenie parametrov trénovania

Na trénovanie sa neodporúča vyberať presvietené snímky a strany, ktoré obsahujú ťažko čitateľné miesta, škrtance, rukopisné poznámky a podobne. Vybrané reprezentatívne strany sa následne rozdelia do cvičného (*training set*) a overovacieho súboru (*validation set*), ideálne v pomere 10 : 1. Softvér ponúka možnosť manuálneho aj automatického výberu strán do trénovania. Ďalším krokom je nastavenie parametrov, ktoré môžu zlepšiť ukazovatele vytrénovaného modelu. Prednastavený maximálny počet cyklov (opakovaní trénovania) je 250 a odporúčame ho ponechať. Zvyšovaním počtu cyklov sa proces trénovania predlžuje a zvyšovanie nemusí mať pozitívny vplyv na výslednú úspešnosť modelu. Voľba predčasného zastavenia trénovania (*Early Stopping*) je predvolená na 20 cyklov, čo znamená, že ak sa hodnoty modelu zlepšujú, trénovanie bude aj po dosiahnutí 20 cyklov pokračovať. Ak však už hodnoty nevykazujú zlepšenie, trénovanie sa automaticky zastaví a vyhodnotí. Funkcia obráteného textu (*Reverse Text*) sa používa v prípadoch, keď je smer písania na snímke opačný ako pri prepise (napr. originál je napísaný sprava doľava a prepísaný text smeruje zľava doprava). Pri trénovaní je možné predvolene použiť polygóny, ktoré boli vygenerované pri segmentácii dokumentu (*Use existing line polygons for training*). V prípade zrušenia tejto voľby sa pri trénovaní modelu vytvoria nové. Voľba trénovania skratiek (*Train Abbrews with expansion*) pomáha dosiahnuť lepšie výsledky pri rozpoznávaní skratiek. Ak sa na stranách

vybraných do vzorky *Ground Truth* nachádzajú slová označené tagmi Nejasný (*unclear*) alebo Medzera (*gap*), z procesu trénovania ich možno vylúčiť voľbou *Omit lines by tag*. Vynechá sa tak nielen označené slovo, ale aj celý riadok, keďže trénovanie prebieha na úrovni riadkov.



Obrázok 3. Odporúčaný počet prepísaných slov vo vzorke GT podľa typu dokumentu a akceptovateľná miera chybovosti znakov po vytrénovaní modelu

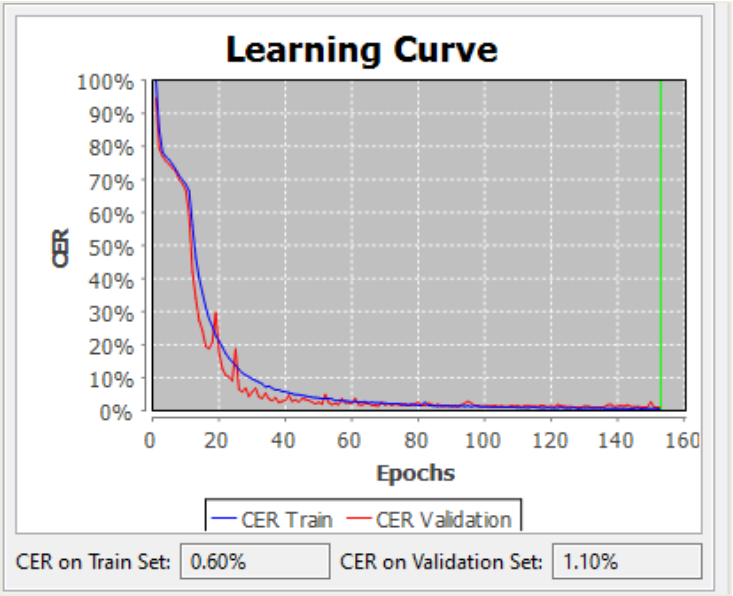
Hodnotenie úspešnosti modelu

Po vytrénovaní modelu expert klient ponúkne vyhodnotenie v podobe grafu a percentuálneho vyjadrenia chybovosti znakov (os y)[27] v automaticky prepísanom texte. Výsledok trénovania CER on Train Set 0,60 % a CER on Validation Set 1,10 % znamená, že v cvičnom súbore bolo bezchybne určených až 99,40 % znakov a v overovacom súbore 98,90 %. Predchádzajúce výskumy v rámci projektu READ ukazujú, že hodnoty CER automaticky prepísaných rukopisných strán do 10 % sa považujú za uspokojivé, v najlepších prípadoch sa pohybujú okolo 5 %. Vynikajúce výsledky modelov trénovaných na tlačených stranách dosahujú chybovosť približne 1 – 2 %[28], preto môžeme výsledky modelu na obr. 4 považovať za veľmi uspokojivé. Druhá časť vyhodnotenia trénovania modelu obsahuje grafické znázornenie procesu trénovania, tzv. krivku učenia sa

[27] Miera chybovosti znakov CER (*Character Error Rate*) sa udáva v percentách. Krivka sa vždy začína na 100 % a tým, ako sa model trénuje a zlepšuje, postupne klesá. Miera chybovosti znakov porovnáva celkový počet znakov (n) vrátane medzier s minimálnym počtom vložení (i), nahradení (s) a vymazaní (d) znakov potrebných na dosiahnutie rovnakého výsledku ako vo vzorke *Ground Truth*. Vzorec na výpočet miery chybovosti znakov: CER = [(i + s + d) / n]*100

[28] MÜHLBERGER, Günter – SEAWARD, Louisse – TERRAS, Melissa et al. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: *Journal of Documentation* [online]. 2019, vol. 75, no. 5, pp. 954–976 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.1108/JD-07-2018-0114

(*Learning Curve*). Po spustení tréningového procesu sa model opakovane trénuje a overuje. V prípade modelu na obr. 4 sme ponechali nastavenie na 250 cyklov (*Epochs*), pričom model sa prestal zlepšovať pri 153. cykle (os x) a trénovanie sa automaticky zastavilo.



Obrázok 4. Výsledok trénovania modelu – grafické znázornenie. Zdroj: Transkribus

Úspešnosť modelu možno vyhodnotiť aj na úrovni jednotlivých strán. Komparáciou textových verzií pomocou nástroja *Compare Text Versions* získame podrobný prehľad toho, čo model prepísal správne a kde v porovnaní s verziou *Ground Truth* urobil chybu.


Version Comparator

Show line numbers

~~K a m e n j.~~ Kamenj.
Pjſek 1. a ~~Kyſy~~ Kysy (Kamenčj) 2. geſt roztroſkotaný (ſetřený,
~~ſtlučený~~ ſtlučený, zdrobený) Kámen (Kamýnek.)
Kamen 3. geſt ~~Cáſt (Cáſtka) Skály.~~ Část (Částka) Skály. 4.
~~Bruo,~~ Brus, 5. Sſkřemen (Křemen,) 6. Mramor, 7. gſau ~~temnė~~ temné
(tmawé, zatmělé) Kamenj.
~~Magnet~~ Magnet 8. přitahuge Železo k ſobě.
Drahé Kamenj 9. gſau ~~Kamýnkowe průhledčiwj,~~ Kamýnkowé průhledřiwj, (ſwětlj:)
gako: bjlý Diamant ~~(Dėmant), čerwený~~ (Démant), čerwený Rubjn, modrý
~~Zaſjr,~~ Zafjr, zelený ~~Sſmaragd,~~ Sſmaragd, žlutý Jacynt. a. g. p.
a bliſſtj ~~ſe~~ ſe, kdyż gſau vhlowatj.
~~Drobne~~ Drobné Perly a weliké Perly 10. roſtau w ~~Oſtrigách~~ Oſtrigách (moř-
~~ſtých Hlemeyždjch,~~ ſkých Hlemeyždjch, Sſkořepinách.)
Korále ~~11.~~ 11 w Moři na Podobu ~~(Způsob)~~ (Způſob) malých Stromkůw.
~~Kůtſſtayn~~ Agtſſtayn 12. zbjrá ſe při Moři, zwlášſſ w Průſých (w
~~Průoſké~~ Průsſké Zemi.)
~~Sſo~~ Sklo 13. ~~podobnė~~ podobné geſt Kriſſtálu.

Obrázok 5. Nesprávne prepísané slová v overovacom súbore (červené) porovnané so správnym prepisom vo vzorke Ground Truth (zelené). Zdroj: Transkribus

Nástroj *Compare* umožňuje v percentách vyjadriť mieru chybovosti na úrovni slov WER (*Word Error Rate*). Tá sa môže na jednotlivých stranách výrazne líšiť, treba si však uvedomiť, že zväčša ide o chyby spojené s interpunkciou (chýbajúca alebo nadbytočná bodka, čiarka, dvojbodka a pod.) a s diakritikou (krátka samohláska namiesto dlhej, spoluhláska namiesto spoluhlásky s mäkčeňom a pod.), ktoré majú minimálny alebo žiadny vplyv na zrozumiteľnosť textu.

V texte získanom po automatickej transkripcii a nevyhnutných korektúrach je možné zvýrazniť dôležité údaje a označiť jednotlivé časti štruktúry dokumentu. Platforma Transkribus ponúka nástroje na prácu s textom prostredníctvom tagovania (značkovania). K dispozícii sú dva druhy tagov. Textové tagy zdôrazňujú významné informácie v texte, napr. osobné mená, geografické názvy, dátumy, skratky a pod. Štrukturálne tagy definujú jednotlivé časti štruktúry dokumentu ako nadpisy, odseky, ilustrácie, čísla strán, marginálne poznámky a i.[29] Okrem exportu

[29] BÔBOVÁ, Mária – KATUŠČÁK, Dušan – KURHAJCOVÁ, Alica et al. *Automatická transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus: Metodická príručka pre účastníkov workshopu* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2024, s. 113-122 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2024.9788055721439

snímok a prepísaného textu je z platformy možné exportovať aj textové tagy využiteľné na tvorbu registrov.

Trénovanie vlastných modelov

Trénovanie samostatných modelov pre jednotlivé fonty skúmanej tlače pomocou technológie HTR+ v prvej fáze projektu neprinieslo výsledky, ktoré by zodpovedali vývojármi deklarovanej nízkej chybovosti pri tlačených dokumentoch. V ukazovateľoch CER a WER na overovacom súbore a aj vizuálne (porovnávaním textových verzií) v spoločných modeloch pre všetky fonty modely vykazovali vysokú chybovosť vo fonte švabach. Už spomínané technologické zmeny v roku 2022 priniesli odstavenie dosiaľ používanej technológie HTR+ a nasadenie nového stroja PyLaia. Z tohto dôvodu sme s časovým odstupom 1,5 roka samostatné modely pretrénovali novou technológiou. Tá mala spočiatku výrazné nedostatky, ktoré sa prejavovali najmä vyššou chybovosťou. Ako najproblematickejší font sa na základe ukazovateľov CER a WER na overovacom súbore ukázala kurzíva a s fontom švabach si technológia neporadila ani po opakovanom spustení trénovania.

Zdokonaľovanie technológie na automatický prepis historických dokumentov, nové funkcionality a nástroje stále priamejšie smerujú k snahe vytvárať univerzálne modely, ktoré budú schopné automaticky prepisovať typovo analogické rukopisné, tlačené či strojové dokumenty. Súbežne so samostatnými modelmi sme sa preto rozhodli trénovať aj spoločné modely, ktoré nezohľadňovali typ fontu, ani jazyk. Dokument, ktorý bol predmetom nášho výskumu, sme nakoniec automaticky transkribovali dovtedy najlepším modelom Model 10 a výsledný text podrobili redakčným úpravám. Relatívne malý rozsah – 184 strán textu formátu A5 – nám umožnil detailne analyzovať chybovosť až na úroveň diakritiky a interpunkcie. V zmysle metodiky pre automatizované vyhodnocovanie úspešnosti modelu pomocou ukazovateľa CER platí, že aj malá chyba pri prepise je štatisticky plnohodnotnou chybou. To znamená, že chýbajúcu čiarku, „u“ namiesto „v“, „á“ namiesto „ä“, medzeru navyše alebo veľké písmeno namiesto malého softvér vyhodnotí ako rovnocenné chyby. Rovnaký princíp sme v procese redakčnej úpravy použili pri typológii týchto chýb a ich následnom štatistickom vyhodnotení (Tab. 1). Položka *Nesprávne prepísaná graféma/interpunkcia* znamená nesprávny prepis samotnej grafémy, najčastejšie „v“ namiesto „u“, „l“ alebo „k“ namiesto „ſ“, ale aj nesprávny prepis interpunkčného znamienka (dĺžeň, dvojbodka, tri bodky, apostrof) na inak správne prepísanej graféme, najčastejšie „ö“ namiesto „ő“ alebo „ů“ či „ű“ namiesto „ů“ alebo „ů“. Najvyššia chybovosť na úrovni strany sa vyskytla vo vecnom

registri v maďarčine tlačenej kurzívou. Viacerí riešitelia projektu Skriptor konštatovali problematický prepis číslic v rukopisných textoch[30], chybovosť v prepise číslic sa však v tlači *Orbis Pictus* (1798) vyskytovala ojedinele s výnimkou už spomínaných registrov. Tie naopak vykazovali vysokú chybovosť vo viacerých aspektoch. Dôvodom bol spôsob rozloženia textu, malé medzery medzi oddelenými stĺpcami a vysoký výskyt interpunkčných znamienok (bodky), ktoré boli použité na oddelenie názvov kapitol a stránkovania. Ako najproblematickejšie písmo sa dovtedy dlhodobo ukazoval švabach, ktorý zvyšoval chybovosť čítania znakov aj v spoločných modeloch. Dokazujú to hodnoty namerané pri samostatnom trénovaní fontov pomocou technológie HTR+. Ukazovatele získané po pretrénovaní tých istých súborov technológiou PyLaia však v miere chybovosti „favorizovali" kurzívu. Podrobná analýza chybovosti v tabuľke toto zistenie potvrdzuje.

Tabuľka 1. Typy a počet chýb po automatickej transkripcii textu Modelom 10

Typ chyby	**antikva**	**kurzíva**	**fraktúra**	**švabach**
nesprávne prepísaná graféma / interpunkcia	430	541	386	548
chýbajúca graféma / interpunkcia / medzera	50	66	61	33
graféma / interpunkcia / medzera, kde nemá byť	39	48	43	48
nesprávne prepísaná diakritika (bodka, čiarka, bodkočiarka, dvojbodka, pomlčka, otáznik)	128	48	51	24
chýbajúca diakritika (bodka, čiarka, bodkočiarka, dvojbodka, pomlčka, zátvorka)	65	27	94	20
diakritika, kde nemá byť (pomlčka, bodka, dvojbodka)	70	50	45	43
Správnosť čítania špeciálnych znakov s nízkym výskytom v tréningovom súbore				
ligatúra ae	1			
ligatúra oe	15			
ligatúra tz			4	
Celkový počet chýb	**798**	**780**	**684**	**716**

Domnievame sa, že najčastejším dôvodom nesprávneho prepisu bola nižšia kvalita samotnej tlače (najmä slabo vytlačené písmo), presvit z opačnej strany listu, otlačenie tlačiarenskej farby z predchádzajúcej strany a nečistota na papieri.

Na zvyšovanie kvality modelov boli v rôznych fázach projektu použité tri základné postupy a ich kombinácie:

- oprava manuálnej transkripcie – po vytrénovaní modelu a analýze chýb na úrovni znakov a slov softvér upozorní aj na vlastné chyby výskumníka. V tejto fáze je teda priestor na opätovnú kontrolu prepisu vzorky GT a opravu nezrovnalostí.
- zvýšenie počtu slov v cvičnom súbore – v niektorých prípadoch je odporúčaný počet slov zahrnutých do trénovania modelu automatickej transkripcie nedostatočný a vtedy je na mieste zvýšiť počet strán.

[30] NAGY, Imrich. Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*. 2021, roč. 51, č. 2, s. 53-67 a TOMEČEK, Oto. O možnostiach automatického prepisu historických rukopisných dokumentov pomocou platformy Transkribus. In: *Historica. Revue pro historii a příbuzné vědy* [online]. 2025, roč. 16, č. 1 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.15452/Historica.2025.16.0005

Vo väčšine prípadov platí pravidlo, že čím vyšší je výskyt znaku v cvičnom súbore, tým lepšie sa ho stroj naučí rozpoznávať. Touto metódou sa dá znížiť miera chybovosti znakov v overovacom súbore aj o niekoľko percentuálnych bodov.

- použitie základného modelu (*base model*) – táto metóda vychádza z predpokladu, že základný model si pamätá, čo sa naučil, a teda každé ďalšie trénovanie ho (teoreticky) zlepšuje. Pri tomto postupe možno ako *base model* použiť najlepší vlastný model alebo zodpovedajúci model iného používateľa.

Z hľadiska ďalšej využiteľnosti ktoréhokoľvek modelu je dôležité, aby bol proces jeho vzniku transparentne popísaný. Na tento účel platforma Transkribus ponúka možnosť vytvoriť redakčné vyhlásenie (*Editorial Declaration*). Táto funkcia obsahuje súbor preddefinovaných položiek 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 a rozhodovanie o jeho použití na prepis iných obsahovo, či formálne príbuzných dokumentov, alebo ako *base modelu*. Je vhodné, ak je redakčné vyhlásenie dostupné aj v anglickom jazyku, keďže komunita používateľov platformy Transkribus je medzinárodná a model môže mať potenciál byť aplikovateľný aj na iné tlače nielen územne, jazykovo a provenienčne slovacikálneho charakteru.

Jedným z najočakávanejších okamihov našej niekoľkoročnej práce bolo použitie najlepšieho vytrénovaného modelu na automatickú transkripciu formálne, obsahovo a typovo porovnateľného dokumentu s cieľom overiť použiteľnosť, resp. posúdiť mieru univerzálnosti tohto modelu. Na tento účel sme zvolili Model 11 s ukazovateľmi CER 0,60 % na cvičnom a CER 1,00 % na overovacom súbore. Modelom sme automaticky prepísali vybrané strany učebnice J. A. Komenského *Orbis Pictus* vydanej v tej istej tlačiarni v roku 1820. K dispozícii sme mali dve vzorky digitalizátov v rôznej kvalite. Prvá vzorka zdigitalizovaná v kvalite 600 DPI v Centre vedecko-technických informácií SR v Bratislave bola v minulosti reštaurovaná, preto v dôsledku pevnosti väzby nebolo možné nasnímať všetky strany bez viditeľného ohybu textu na vnútornom okraji strán. Druhá vzorka zdigitalizovaná v kvalite 96 DPI (dostupná na portáli DIKDA) obsahovala snímku každej strany jednotlivo a bez ohybu v dôsledku väzby. Dve rôzne vzorky nám tak okrem overenia funkčnosti vytrénovaného modelu na podobnej historickej tlači umožnili preveriť aj vplyv kvality digitalizátu na automatickú transkripciu. Na overenie oboch hypotéz sme si vybrali dve strany obsahujúce ten istý text skladajúci sa zo všetkých štyroch fontov – antikva, kurzíva, fraktúra

a švabach. Funkčnosť modelu sa nám na oboch vzorkách potvrdila. Model 11 použitý na pôvodnej tlači *Orbis Pictus* (1798) vykazoval chybovosť na úrovni CER 2,55 % a WER 7,50 %. Na stranách *Orbis Pictus* z roku 1820 Model 11 vykazoval vyššiu chybovosť, ale na tú zrejme mala vplyv aj viditeľne horšia kvalita tlače originálneho diela. Vzorka zdigitalizovaná v kvalite 96 DPI vykazovala chybovosť CER 3,80 % a WER 12,61 %, vzorka zdigitalizovaná v kvalite 600 DPI vykazovala chybovosť CER 4,98 % a WER 16,33 %. Na základe výsledkov možno konštatovať, že kvalita digitalizátu (počet DPI) nemá dramatický vplyv na presnosť prepisu. Nepotvrdili sa tak závery, ktoré vo svojom článku publikovali Harish a Raghavendra[31]. Väčší počet chýb v prípade kvalitnejšieho digitalizátu má za dôsledok mierny ohyb strán v mieste väzby. Vyššia hodnota WER nespochybňuje efektívnosť a účelnosť automatickej transkripcie, keďže ide o nezávažné chyby v diakritike (chýbajúci/nadbytočný dĺžeň a i.) a interpunkcii (rozdeľovník na konci slova, čiarka namiesto bodkočiarky a pod.), ktoré nemajú žiadny vplyv na zrozumiteľnosť textu.

Model 11 sme v ďalšom kroku použili na prepis vybraných strán historického diela Mateja Bela *Adparatus ad Historiam Hungariae*[32] z roku 1735. Ide o sedemstodvadsaťosemstranové dielo polyhistora Mateja Bela napísané v latinskom jazyku, ktoré je tiež súčasťou historického knižničného fondu Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici. Textová časť obsahuje len dva fonty – antikvu a kurzívu. Formálne usporiadanie textu pozostáva z nadpisov, ozdobných iniciálok, marginálií a poznámok pod čiarou. Niektoré časti diela sú usporiadané do stĺpcov a tabuliek. Ozdobnú časť tvoria okrem iniciálok aj vinety a vlysy. V texte sa nachádzajú špecifické ligatúry (ct, ſt, Æ, &, æ), grafémy (ê, ė, è, ù, â, à, á, ù, û, m̃, W) a znaky (§, =, ¿, „). Historický dokument bol zdigitalizovaný na profesionálnom skeneri v Centre vedecko-technických informácií SR v Bratislave v rozlíšení 600 DPI.

Vzhľadom na príbuzné fonty písma, geografickú blízkosť tlačiarov a obdobie vydania diel *Orbis Pictus* (1798) a *Adparatus ad Historiam Hungariae* (1735) autorky predpokladali dobré výsledky automatickej transkripcie s použitím Modelu 11. Výsledný prepis však tieto očakávania

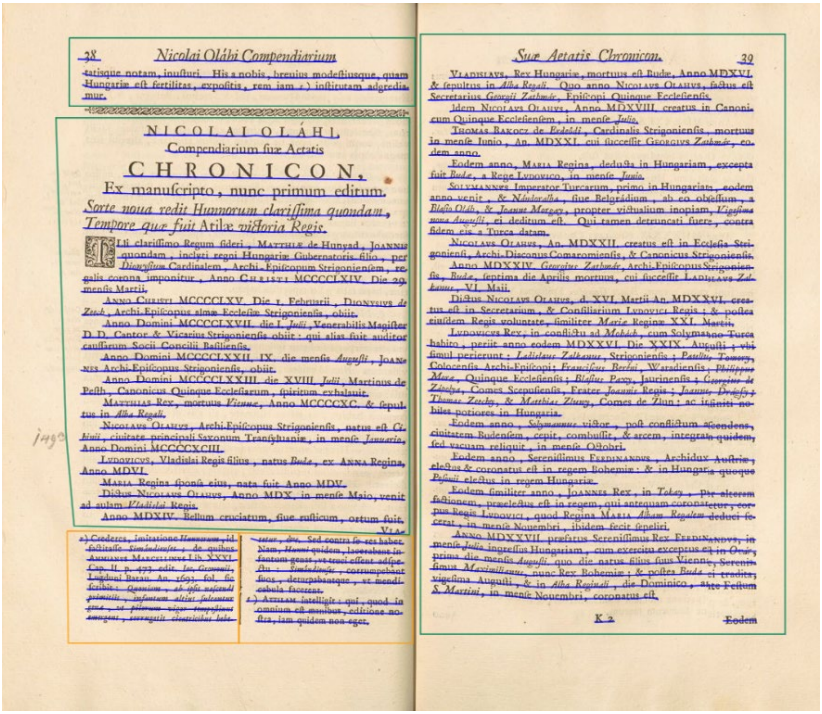
[31] HARISH, R. – RAGHAVENDRA RAO, G. N. Transcription of Ancient Indian Manuscripts Through Artificial Intelligence – Current Status of Technology and the Way Forward. In: *Artificial Intelligence: Theory and Applications, Vol. 844* [online]. Singapore: Springer, 2024, pp. 339-352 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.1007/978-981-99-8479-4_25

[32] BEL, Matthias. *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, 1735.

nepotvrdil. Za dôvody vysokej chybovosti na *Adparate* možno považovať:

- nové typy špecifických grafém, na ktorých nebol Model 11 trénovaný,
- vyšší výskyt kapitálok, ktoré dielo *Orbis Pictus* neobsahovalo v dostatočnej miere.

Významným faktorom zlyhania Modelu 11 sa javí aj kvalita papiera a použitie viacerých druhov antikvy. Na zlepšenie prepisu sme sa preto rozhodli vytrénovať nový model. Príprava vzorky GT vychádzala z opravy chýb na stranách prepísaných Modelom 11[33]. Následne autorky vykonali kontrolu a dodatočné opravy prepisu, čím bola vzorka GT pripravená na trénovanie špecifického modelu Bel_Adparatus_1735_model 1. Model 11 bol pri trénovaní použitý ako *base model*. Nejasné znaky označené ako *unclear* boli z trénovania modelu vylúčené.



Obrázok 6. Ukážka segmentácie tlače Adparatus. Zdroj: Transkribus

Výsledky vytrénovaného modelu boli uspokojivé, poukázali však

[33] Opravu strán zabezpečili študenti Mário Kmecz a Dominik Polička z Katedry filozofie Filozofickej fakulty Univerzity Mateja Bela v Banskej Bystrici v rámci povinnej praxe v Univerzitnej knižnici UMB, ktorým týmto ďakujeme za spoluprácu.

na niekoľko chýb v príprave vzorky GT v podobe nesprávneho prepisu[34]. Po ich odstránení bol vytrénovaný Bel_Adparatus_1735_model 2. V modeli Bel_Adparatus_1735_model 3 sme opätovne otestovali teóriu o vylúčení existujúcich polygónov. Tretí model na tlači *Adparatus ad Historiam Hungariae* nakoniec dosiahol chybovosť len na úrovni 0,93 %.

Univerzálnosť tohto modelu sme recipročne overili na tlači *Orbis Pictus* z roku 1820. Podobne ako v prípade overovania funkčnosti Modelu 11 sme zvolili tie isté strany obsahujúce všetky typy fontov. Výsledky boli neuspokojivé, dosahovali chybovosť na úrovni CER 23,99 % a WER 53,55 % v prípade digitalizátov v kvalite 96 DPI a CER 26,99 % a WER 57,5 % na digitalizátoch v kvalite 600 DPI. Vyššia chybovosť bola spôsobená zlou čitateľnosťou textu v blízkosti väzby dokumentu. Model Bel_Adparatus_1735_model 3 mal problém s čítaním špecifických grafém s dnes už nepoužívanou interpunkciou (ő, ű, ċ, ė, ṅ, ṙ), ktorá sa v tlači *Adparatus ad Historiam Hungariae* nevyskytovala. Keďže Bel_Adparatus_1735_model 3 bol vytrénovaný na fontoch antikva a kurzíva, predpokladali sme dobrý výsledok transkripcie v týchto fontoch. V kurzíve však bola chybovosť na zvlášť vysokej úrovni, a to nielen pri prepise špecifických grafém. Chybovosť vo fontoch fraktúra a švabach nebola až taká prekvapivá. Použitím Modelu 11 ako *base modelu* sa teda nevytvoril univerzálny model vhodný na prepis oboch tlačí. Prehľad a popis všetkých vytrénovaných modelov sa nachádza v prílohe tohto článku.

V roku 2023 platforma Transkribus už obsahovala také množstvo dát, že vývojárom poslúžili ako základ na vytrénovanie prvých takzvaných supermodelov. Tie sú založené na výkonnejšej technológii a trénované na veľkom objeme rôznorodých dokumentov. Supermodely výrazne uľahčujú využívanie platformy, pretože (teoreticky) umožňujú transkripciu veľkých zbierok obsahujúcich rôzne jazyky a viac typov písma bez nevyhnutnosti vytvárania vlastného modelu[35]. Začiatkom roka 2024 na výročnej konferencii Transkribus komunity TUC24 vývojový

[34] Skúsenosti s poukázaním na vlastné chyby pri príprave súboru GT sa potvrdili pri trénovaní modelov pre tlače *Orbis Pictus* aj *Adparatus ad Historiam Hungariae*. Na rovnaké skúsenosti poukázali aj ďalší autori, napr. GRIFFITHS, Rachael M. Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script. In: *Revue d'Etudes Tibétaines* [online]. 2024, no. 72, pp. 43-51 [cit. 2026-03-26]. Dostupné na: https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_72_03.pdf; RAGHALLAIGH, Brian. O. – PALANDRI, Andrea – CÁRTHAIGH, Críostóir Mac Handwritten Text Recognition (HTR) for Irish-Language Folklore. In: *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022* [online]. Paris: European Language Resources Association, 2022, pp. 121-126 [cit. 2026-03-26]. Dostupné na: https://aclanthology.org/2022.cltw-1.17

[35] PARK, Fiona. What are Super Models and how do they work? In: *Blog.transkribus.org* [online]. April 22, 2025 [cit. 2026-03-26]. Dostupné na: https://blog.transkribus.org/en/what-are-super-models-and-how-do-they-work

tím avizoval aplikovanie veľkých jazykových modelov (Large Language Models – LLM) do platformy. Postupné zdokonaľovanie nástrojov umelej inteligencie využívajúcich veľké jazykové modely umožnilo dosahovať veľmi dobré výsledky prepisu rukopisných dokumentov aj bez špecificky vytrénovaných modelov[36]. Výsledky testovania supermodelu *Text Titan I ter* vykazujú jeho vysokú univerzálnosť a zároveň vyššiu presnosť v porovnaní s predchádzajúcimi LLM nástrojmi ako ChatGPT, Gemini, Claude a Mistral[37]. Aj výsledky práce na úlohách projektu Skriptor a vytrénované modely viedli k vytvoreniu dvoch agregovaných modelov – *Slovak Supermodel M1 (SSM1)* pre rukopisné texty a *Slovak Supermodel print&typewriter1 (SSPT1)* pre historické tlače a strojopisný text[38]. Dokumenty a dáta spracované riešiteľmi boli s veľkou pravdepodobnosťou použité aj na trénovanie jedného z prvých supermodelov *Transkribus Print M1*, ktorý obsahuje aj tlače v slovenskom a českom jazyku.

Supermodel Transkribus Print M1 sme otestovali na automatickom prepise tlače *Orbis Pictus* (1820). Výsledky transkripcie boli na vysokej úrovni. Prepis bol čitateľný vo všetkých štyroch fontoch, chyby sa vyskytovali len v prípade diakritiky a interpunkcie s minimom chýb v prepise grafém a číslic. Transkripcia však neprihliadala na špecifické, dnes už nepoužívané grafémy použité predovšetkým vo fontoch kurzíva a švabach (napr. ő, ű, ċ, ė, ṅ, ṙ). Okrem toho je zaujímavé, že supermodel bol pravdepodobne vytrénovaný na vzorke, ktorá zohľadňovala fonetické čítanie jednotlivých slov a grafému g automaticky prepisovala na grafému j (gsou → jsou, gako → jako, gezerné → jezerné, plauge → plauje a i.). Pre používateľov platformy Transkribus, ktorí netrvajú na prepise formou transliterácie, je tak použitie tohto modelu jasnou voľbou. Pre potreby nášho projektu by prepis dokumentu *Orbis Pictus* v kvalite 96 DPI s použitím tohto modelu a následnej opravy špecifickej interpunkcie vykazoval chybovosť CER 7,78 % a WER 21,97 %. Ak sa používateľ platformy uspokojí s transkripciou, chybovosť modelu je CER 3,06 % a WER

[36] HUMPHRIES, Marc – LEDDY, Lianne C – DOWNTON, Quinn et al. Unlocking the Archives: Large Language Models Achieve State-of-the-Art Performance on the Transcription of Handwritten Historical Documents. In: *ArXiv* [online]. 2024 [cit. 2026-03-26]. Dostupné na: https://arxiv.org/html/2411.03340v1

[37] NOBLE, Aurelius. The new Text Titan I ter and how it compares to ChatGPT, Gemini, and other LLMs. In: *Blog.transkribus.org* [online]. June 20, 2025 [cit. 2026-03-26]. Dostupné na: https://blog.transkribus.org/en/new-text-titan-i-ter-and-how-it-compares-to-chatgpt-gemini-and-other-llms

[38] BÔBOVÁ, Mária – KATUŠČÁK, Dušan – KURHAJCOVÁ, Alica et al. *Automatická transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus: Metodická príručka pre účastníkov workshopu* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2024, s. 100-103 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2024.9788055721439

10,56 %. Do tejto chybovosti je započítaný aj prepis grafémy g, ktorú model prepisoval ako j.

Závery

Štvorročná práca s platformou Transkribus dáva priestor na formulovanie nasledujúcich záverov. Manuálna segmentácia celého dokumentu a manuálny prepis (alebo poloautomatický prepis s následnou korektúrou) vybraných strán, ktoré sa použijú ako vzorka *Ground Truth* na vytrénovanie nového modelu, sú najzdĺhavejšou časťou prípravných prác predchádzajúcich automatickej transkripcii celého dokumentu. S pribúdajúcimi skúsenosťami s prácou na platforme sa však časová dotácia na tieto činnosti znižuje v prospech výskumníka.

Trénovaniu modelu výrazne pomáha vynechanie problematických strán (preexponované snímky, nečistoty na papieri, machule) a ťažko čitateľných miest v texte (škrtance, zhustené písmo na konci riadkov, zahnuté písmo v strede pri väzbe, slabá tlač). Vidieť to najmä na CER overovacieho súboru, kde rozdiely na jednotlivých stranách dosiahli viac ako 1 %, a na ukazovateli WER, kde rozdiel chybovosti na dvoch overovacích stranách toho istého modelu dosiahol aj viac ako 10 %. K rovnakému záveru dospeli aj riešitelia, ktorí skúmali rukopisné dokumenty[39].

Navýšenie strán vo vzorke *Ground Truth* oproti odporúčanému rozsahu nemá výrazný vplyv na zlepšenie modelu. Platí to pre rukopisné aj tlačené dokumenty. Dôkazom sú modely 10 a 11, ktoré sme trénovali s odstupom takmer jedného roka, pričom Model 11 bol trénovaný na celej tlači s rozsahom 184 strán, avšak na ukazovateľoch vidieť len minimálne zlepšenie. Navyšovanie má naopak zmysel v prípade, že sa v texte (a teda aj v cvičnom súbore) v menšom počte nachádzajú špeciálne znaky a ligatúry. Znak by sa mal v texte vyskytovať v dostatočnom počte (aspoň 50x pri tlačiach a 500x pri rukopisoch), aby sa ho stroj naučil (spoľahlivo) čítať.

Vo všetkých prípadoch trénovania modelu sa dosiahli lepšie výsledky s predvoľbou trénovania bez existujúcich polygónov. Tento záver platí aj pre trénovanie modelov na tlači *Adparatus ad Historiam Hungariae*. Na rovnakých cvičných a overovacích stranách dosahoval Bel_Adparatus_1735_model 3 trénovaný bez použitia existujúcich polygónov lepšie výsledky ako Bel_Adparatus_1798_model 2 (na úrovni chybovosti znakov

[39] NAGY, Imrich. Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. In: *Slovenská archivistika*. 2021, roč. 51, č. 2, s. 53-67 a BÔBOVÁ, Mária – KATUŠČÁK, Dušan – KURHAJCOVÁ, Alica et al. *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2022 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2022.9788055720203

CER ide o rozdiel 0,30 %).

Pri trénovaní modelov pomocou PyLaia je automaticky prednastavená voľba *Early stopping* na 20 cyklov. Odporúčame toto nastavenie ponechať, pretože ak sa výsledky nezlepšujú, softvér pokračuje v trénovaní modelu. Dokazuje to nastavenie parametrov pri trénovaní vybraných modelov. Model Bel_Adparatus_1735_model 2 nechal Transkribus trénovať 250 cyklov aj napriek tomu, že krivka výsledkov na úrovni chybovosti znakov (CER) sa približne od 65 cyklu nezlepšovala. V prípade modelu Bel_Adparatus_1735_model 3 sa krivka v priebehu 250 cyklov trénovania menila. Model Bel_Adparatus_1735_model 3 dosahoval o 0,30 % lepšie výsledky v CER. Pri modeloch trénovaných na tlači *Orbis Pictus* (1798) sa v jednom prípade zastavilo trénovanie po 153 cykloch (Model 10), v ďalšom po 92 cykloch (Model 11). Rozdiel na úrovni CER bol iba 0,10 % a to aj napriek tomu, že Model 11 obsahoval podstatne rozsiahlejšiu vzorku cvičných a overovacích dát.

Jednoznačne sa ukazuje, že je prínosom aplikovať *base model* pri trénovaní modelov na tom istom dokumente, či už zodpovedajúci model iného používateľa alebo vlastný model. Vidieť to na grafoch modelov Bel_Adparatus_1735_model 1 a 2. Transkribus *s base modelom* vykázal na krivke učenia už pri prvom cykle nižšiu neúspešnosť (v rozmedzí 87 – 95 % CER na overovacom súbore), pričom bez použitého *base modelu* je neúspešnosť pri prvom cykle takmer vždy 100 %.

Zdokonalenie modelu jeho obohatením o ďalší typ písma prostredníctvom *base modelu* trénovanom na inom dokumente sa nepotvrdilo. Model Bel_Adparatus_1735_model 3 vytrénovaný na tlači *Adparatus ad Historiam Hungariae* sme v rámci overovania jeho funkčnosti použili na vzorke jednej dvojstrany tlače *Orbis Pictus* (1820). Výsledky boli neuspokojivé. Napriek tomu, že pri jeho trénovaní bol ako *base model* použitý Model 11, nový model nevykazoval dobrú čitateľnosť špecifických znakov a fontov fraktúra a švabach.

V súvislosti s AI sa veľa hovorí a píše o jej nestabilite a náchylnosti halucinovať (vytvárať nesprávne alebo zavádzajúce výsledky). Faktom však je, že moderné modely umelej inteligencie, najmä veľké jazykové modely (LLM), sú navrhnuté tak, aby rozpoznávali vzory v rozsiahlych súboroch údajov a generovali vierohodné odpovede na základe týchto vzorov. Z tohto dôvodu sme sa rozhodli najúspešnejší model Bel_Adparatus_1735_model 3 preveriť jeho opakovaným trénovaním s tými istými parametrami. Výsledok svedčí o vysokej konzistentnosti Transkribusu – model Bel_Adparatus_1735_model 3a sa v ukazovateli CER mierne zlepšil (0,27 % oproti 0,31 %), ukazovateľ WER zostal rovnaký.

Textové verzie (DOCx, PDF, Excel) získané „z obrázkov" automatickou transkripciou sú fulltextovo prehľadávateľné a vďaka rozvoju ďalších nástrojov umelej inteligencie na automatizovaný preklad je možné využiť prekladače z rôznych, aj menej používaných jazykov. Takto transkribovaný a preložený dokument môže slúžiť na ďalší vedecký výskum a jeho spracovanie v podobe pramennej edície [40].

Využitie platformy má zmysel predovšetkým pri práci s rozsiahlymi dokumentmi alebo zbierkami. V prípade dokumentov s menším počtom strán nemusí pomer času a práce na vytrénovaní nového modelu na jednej strane a výsledný automaticky transkribovaný text s väčšou či menšou mierou chybovosti na druhej strane vychádzať v prospech úsilia výskumníka. V tomto subjektívnom hodnotení vychádzame z toho, že s expert klientom sme sa od začiatku projektu učili pracovať metódou pokus – omyl, samoštúdiom vtedy relatívne obmedzeného počtu metodík dostupných len na webových stránkach projektu READ a občasnou výmenou skúseností medzi členmi riešiteľského kolektívu. Motivačne nepôsobilo ani to, že uprostred projektu vývojový tím odstavil dosiaľ používanú technológiu a dovtedajšie výsledky trénovania modelov v podstate zostali bezcenné, resp. nepoužiteľné, ak sme chceli porovnávať výsledky s modelmi, ktoré sme v nasledujúcom období trénovali novou technológiou. Napriek tomu rýchly vývoj platformy postupne prinášal zlepšenia, ktoré by nám počiatočné práce na segmentácii a príprave GT vzorky výrazne zjednodušili a ušetrili množstvo času. Zároveň však treba zdôrazniť, že získané skúsenosti priniesli benefit vo forme vynikajúceho modelu na prepis takmer osemstostranového diela Mateja Bela a jeho príprava a trénovanie si s odstupom takmer štyroch rokov vyžiadali približne desatinu času a úsilia v porovnaní s prácami na Komenského učebnici *Orbis Pictus*.

Na základe skúseností s platformou Transkribus a vyvodených záverov odporúčame jej využívanie na automatickú transkripciu historických dokumentov. Kontinuálny vývoj, zlepšovanie a pridávanie funkcionalít, ako aj rastúci záujem používateľov sú dôkazom, že medzi nástrojmi na automatický prepis starých rukopisov a tlačí má svoje významné miesto. V súčasnosti je Transkribus najčastejšie používaným nástrojom HTR v oblasti kultúrneho dedičstva [41]. Široké využitie platformy sa ukazuje

[40] TOMEČEK, Oto – NAGY, Imrich. *Reambulačný protokol mesta Banská Bystrica z roku 1820: edícia prameňa pomocou nástroja na automatickú transkripciu historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2024 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.24040/2024.9788055721903

[41] NOCKELS, Joe – GOODING, Paul – AMES, Sarah – TERRAS, Melissa. Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research. In: *Archival Science* [online]. 2022, vol. 22, no. 3, pp. 367-392 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.1007/s10502-022-

nielen v oblasti trénovania modelov, ale aj rozpoznávania typov rukopisov[42], či trénovania modelov na automatické rozpoznávanie entít nachádzajúcich sa v dobových textoch[43]. Platforma dnes poskytuje používateľom širokú podporu: organizuje webináre, semináre a workshopy na rôzne témy nielen v angličtine, ale aj v iných jazykoch[44], k dispozícii sú manuály a videá na prácu so softvérom, každoročne sa koná výročná konferencia *Transkribus User Conference*. V roku 2020 získal Transkribus cenu *Horizon Impact Award*, ktorá sa udeľuje projektom so spoločenským dopadom na celú Európu i za jej hranicami. V januári 2023[45] komunitu tvorilo viac ako 100 000 individuálnych a inštitucionálnych používateľov (vrátane národných knižníc, akademických knižníc, archívov, múzeí a výskumných inštitúcií[46]) a na serveroch bolo uložených 43 miliónov spracovaných digitalizovaných strán rukopisných a tlačených dokumentov. V júni 2026 počet používateľov prekročil hranicu pol milióna, množstvo prepísaných strán dosiahlo 200 miliónov a počet vytrénovaných HTR AI modelov 20 000.

Napriek výraznému pokroku v oblasti automatickej transkripcie historických dokumentov zostáva otázka univerzálnosti modelov otvorená. Výsledky výskumu ukázali, že aj modely dosahujúce veľmi nízku mieru chybovosti pri konkrétnom type dokumentu nemusia byť bez ďalšieho trénovania spoľahlivo aplikovateľné na formálne alebo jazykovo príbuzné tlače. Ukazuje sa, že limity univerzálnosti sú podmienené najmä kvalitou digitalizátu, variabilitou historických fontov, prítomnosťou špecifických grafém, rozdielnou typografiou dokumentov a rozdielmi v tlačiarenských technikách.

Súčasný vývoj však naznačuje výrazný posun smerom k robustnejším univerzálnym modelom založeným na veľkých dátových súboroch a architektúrach využívajúcich veľké jazykové modely (LLM). Supermodely typu Transkribus Print M1 alebo Text Titan predstavujú prechod od modelov viazaných na konkrétny dokument k modelom schopným efektívne spracúvať rozsiahle heterogénne zbierky historických tlačí

09397-0

[42] CUÉLLAR, Álvaro – BOADAS, Sònia. Artificial Intelligence for Calligraphic Writer Identification: The Case of Lope de Vega's Autographs. In: *Hipogrifo. Revista de literatura y cultura del Siglo de Oro* [online]. 2025, vol. 12, no. 1, pp. 517-532 [cit. 2026-03-26]. Dostupné na: https://doi.org/10.13035/H.2025.13.01.36

[43] SÁNCHEZ-SALIDO, Eva – MENTA, Antonio – GARCÍA-SERRANO, Ana. Seeking information in Spanish Historical Newspapers: The Case of Dario de Madrid (18th and 19th Centuries). In: *Digital Humanities Quarterly* [online]. 2023, vol. 17, no. 4 [cit. 2026-03-26]. Dosutpné na: huttps://dhq.digitalhumanities.org/vol/17/4/000735/000735.html

[44] https://www.transkribus.org/events

[45] https://edinburgh-innovations.ed.ac.uk/case-studies/transforming-scholarship-in-the-archives

[46] https://readcoop.org/members

bez potreby náročného lokálneho trénovania. V budúcnosti možno očakávať ďalšie znižovanie potreby manuálnych zásahov, automatizáciu korektúr a presnejšie zohľadňovanie historických jazykových a typografických špecifík.

Na druhej strane sa zároveň ukazuje, že pri dokumentoch s vysokou mierou grafickej variability, neštandardnou ortografiou alebo pri požiadavke na diplomatický prepis zostane úloha špecializovaných modelov aj naďalej významná. Budúci vývoj preto pravdepodobne nepovedie k úplnému nahradeniu špecificky trénovaných modelov univerzálnymi riešeniami, ale skôr ku kombinácii oboch prístupov – univerzálnych supermodelov na základnú transkripciu a lokálne adaptovaných modelov na vedecké, archívne a filologické účely.

Z pohľadu digitálnych humanitných vied predstavujú univerzálne modely významný potenciál pre masové sprístupňovanie kultúrneho dedičstva, budovanie prehľadávateľných digitálnych korpusov a efektívnejšie prepájanie historických prameňov naprieč pamäťovými inštitúciami. Ich širšie využitie však bude závisieť nielen od technologického pokroku, ale aj od dostupnosti kvalitných tréningových dát, transparentnosti metodiky transkripcie a spolupráce pri tvorbe otvorených dátových a modelových zdrojov.

Budúci vývoj automatickej transkripcie historických dokumentov pravdepodobne povedie k postupnému prepájaniu univerzálnych transkripčných modelov s nástrojmi obsahovej analýzy a veľkými jazykovými modelmi (LLM). Kým súčasná generácia modelov sa primárne zameriava na čo najpresnejšie rozpoznávanie textu, ďalšia etapa vývoja bude smerovať k inteligentnému spracovaniu významu dokumentov, automatickej sumarizácii obsahu, identifikácii entít, tematickému triedeniu a vytváraniu kontextových väzieb medzi historickými prameňmi. Perspektívu takéhoto smerovania potvrdzuje aj projekt DIGINARCH-AI[47], ktorý nadväzuje na výsledky projektu SKRIPTOR a rozširuje ich smerom k automatickej transkripcii rozsiahlych archívnych fondov strojopisných dokumentov a k vývoju inteligentného modelu sumarizácie historických textov. Projekt DIGINARCH-AI zároveň naznačuje posun od experimentálneho overovania modelov k ich implementácii v reálnom archívnom prostredí a k praktickému využívaniu umelej inteligencie pri modernizácii služieb pamäťových inštitúcií. Ambíciou projektu je nielen vytvorenie univerzálneho modelu pre automatickú transkripciu historických strojopisných dokumentov s chybovosťou pod 3 % na úrovni znakov, ale aj

[47] Projekt bol podporený Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-24-0049 *Digitalizácia a inteligentná obsahová analýza historických dokumentov pre odbornú archívnu činnosť a historickú vedu (DIGINARCH-AI)*.

sprístupnenie približne 232 000 strán dokumentov prostredníctvom inteligentných nástrojov vyhľadávania a sumarizácie obsahu.

Takýto vývoj otvára perspektívu budovania rozsiahlych digitálnych korpusov historických dokumentov, ktoré budú nielen strojovo čitateľné, ale aj obsahovo analyzovateľné a používateľsky prístupné širšej odbornej i laickej verejnosti. Univerzálne modely automatickej transkripcie sa tak môžu stať základnou infraštruktúrou digitálnych humanitných vied, archívnictva a knižničnej praxe. Ich širšie uplatnenie však bude aj naďalej závisieť od kvality tréningových dát, dostupnosti otvorených datasetov a schopnosti modelov reflektovať jazykovú, grafickú a typografickú variabilitu historických dokumentov.

Napriek rýchlemu technologickému pokroku preto nemožno očakávať úplné nahradenie odbornej ľudskej kontroly. Skôr sa predpokladá vznik hybridného modelu spolupráce medzi expertom a umelou inteligenciou, v ktorom budú univerzálne modely zabezpečovať rýchlu základnú transkripciu a inteligentné spracovanie obsahu, zatiaľ čo odborník bude garantovať interpretačnú presnosť, kritickú kontrolu a vedeckú validáciu výsledkov.

Príloha

Tabuľka 2. Prehľad samostatných modelov vytrénovaných v rokoch 2022 – 2024

Model	Dátum vytvorenia	ID modelu	Technológia	Cvičný súbor			Overovací súbor					Počet cyklov
				počet riadkov	počet slov	CER	počet riadkov	počet slov	CER	WER 1	WER 2	
Print_antikva_1	19.8.2022	44152	CITlab HTR+	383	917	0,16 %	77	198	2,64 %	8,04 %	5,20 %	50
Print_kurziva_1	19.8.2022	44148	CITlab HTR+	373	1 018	0,13 %	87	236	2,11 %	6,49 %	9,38 %	50
Print_fraktura_1	21.8.2022	44173	CITlab HTR+	370	1 025	0,08 %	88	243	1,23 %	1,94 %	7,18 %	50
Print_schwabach_1	19.8.2022	44150	CITlab HTR+	151	908	0,04 %	31	170	4,78 %	8,74 %	18,71 %	50
Print_antikva_2	12.3.2024	60660	PyLaia	383	917	1,10 %	77	198	2,20 %	8,09 %	6,25 %	201
Print_kurziva_2	12.3.2024	60662	PyLaia	373	1 018	1,20 %	87	236	4,30 %	9,73 %	14,80 %	201
Print_fraktura_2	12.3.2024	60664	PyLaia	370	1 025	0,10 %	88	243	3,00 %	6,45 %	13,88 %	201
Print_schwabach_2	13.3.2024	60669	PyLaia	151	908	99,80 %	31	170	100 %	28,49 %	33,18 %	201

Tabuľka 3 Prehľad spoločných modelov vytrénovaných v rokoch 2022 – 2025[48]

Model	Dátum vytvorenia	ID modelu	Technológia	Cvičný súbor			Overovací súbor					Počet cyklov
				počet riadkov	počet slov	CER	počet riadkov	počet slov	CER	WER 1	WER 2	
Model 1	8.8.2022	43995	CITlab HTR+	653	2 047	0,29 %	202	605	2,33 %	6,42 %	8,56 %	50
Model 2	8.8.2022	44001	CITlab HTR+	601	1 878	0,25 %	254	773	3,09 %	6,65 %	9,74 %	50
Model 3	18.8.2022	44136	CITlab HTR+	653	2 047	0,28 %	202	605	2,23 %	7,74 %	7,31 %	50
Model 4	18.8.2022	44117	CITlab HTR+	601	1878	0,25 %	254	773	2,69 %	6,65 %	9,00 %	50
Model 5	23.8.2022	44220	CITlab HTR+	1 318	4 107	0,36 %	202	603	1,02 %	2,43 %	3,34 %	50
Model 6	23.8.2022	44218	CITlab HTR+	1 266	3 938	0,44 %	254	772	1,53 %	5,08 %	3,72 %	50
Model 7	24.8.2022	44247	CITlab HTR+	653	2 049	0,17 %	202	603	2,26 %	4,87 %	8,35 %	50
Model 8	24.8.2022	44251	CITlab HTR+	601	1 880	0,17 %	254	772	2,64 %	6,85 %	9,14 %	50
Model 9a	5.11.2022	45904	PyLaia	1 881	5 818	1,00 %	243	751	1,20 %	3,69 %	5,12 %	101
Model 9	11.11.2022	46060	PyLaia	1 826	5 791	2,80 %	248	778	2,80 %	12,63 %	7,30 %	101
Model 10	1.4.2023	51087	PyLaia	1 877	5 811	0,60 %	247	760	1,10 %	4,56 %	3,47 %	153
Model 11	1.3.2024	60244	PyLaia	9 854	30 892	0,60 %	1 170	3 653	1,00 %	1,84 %*	5,05 %*	92
Bel_Adparatus_1735_model 1	14.8.2025	385017	PyLaia	1 477	10 337	0,32 %	203	1 626	1,50 %	3,39 %**	8,49 %**	250
Bel_Adparatus_1735_model 2	16.8.2025	386277	PyLaia	1 347	9 411	0,33 %	203	1 623	1,23 %	1,92 %**	8,49 %**	250
Bel_Adparatus_1735_model 3	6.9.2025	395837	PyLaia	1 347	9 411	0,31 %	203	1 623	0,93 %	2,33 %**	6,56 %**	250
Bel_Adparatus_1735_model 3a	23.9.2025	405517	PyLaia	1 347	9 411	0,27 %	203	1 623	0,96 %	2,33 %**	6,95 %**	250

[48] * Do overovacieho súboru modelu 11 bolo zaradených 9 strán (v predchádzajúcich modeloch vždy po 2 strany), WER1 a WER2 Modelu 11 preto uvádzame v tabuľke nasledovne: WER1 = najnižšia hodnota z deviatich, WER2 = najvyššia hodnota z deviatich. ** Do overovacích súborov Bel_Adparatus_1735_model 1, Bel_Adparatus_1735_model 2, Bel_Adparatus_1735_model 3

Model 1: ID43995 MIX print_antikva_kurziva_fraktura_svabach_1 (počet slov: 2047, training set: pages 1 – 8, validation set: pages 9 – 10) – vytvorený 8.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,29 %, bez base modelu, bez existujúcich polygónov

Model 2: ID44001 MIX print_antikva_kurziva_fraktura_svabach_2 (počet slov: 1878, training set: pages 1 – 6, 9 – 10, validation set: pages 7 – 8) – vytvorený 8.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,25 %, bez base modelu, bez existujúcich polygónov

Model 3: ID44136 MIX print_antikva_kurziva_fraktura_svabach_3 (počet slov: 2047, training set: pages 1 – 8, validation set: pages 9 – 10) – vytvorený 18.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,28 %, bez base modelu, bez existujúcich polygónov, po oprave chýb v manuálnom prepise vzorky *Ground Truth*

Model 4: ID44117 MIX print_antikva_kurziva_fraktura_svabach_4 (počet slov: 1878, training set: pages 1 – 6, 9 – 10, validation set: pages 7 – 8) – vytvorený 18.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,25 %, bez base modelu, bez existujúcich polygónov, po oprave chýb v manuálnom prepise vzorky *Ground Truth*

Model 5: ID44220 MIX print_antikva_kurziva_fraktura_svabach_5 (počet slov: 4107 training set: pages 1 – 8, 11 – 15, validation set: pages 9 – 10) – vytvorený 23.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,36 %, bez base modelu, bez existujúcich polygónov, zvýšenie počtu slov v cvičnom súbore

Model 6: ID44218 MIX print_antikva_kurziva_fraktura_svabach_6 (počet slov: 3938, training set: pages 1 – 6, 9 – 15, validation set: pages 7 – 8) – vytvorený 23.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,44 %, bez base modelu, bez existujúcich polygónov, zvýšenie počtu slov v cvičnom súbore

Model 7: ID44247 MIX print_antikva_kurziva_fraktura_svabach_7 (počet slov: 2049 training set: pages 1 – 8, validation set: pages 9 – 10) – vytvorený 24.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,17 %, s base modelom ID43995 MIX print_antikva_kurziva_fraktura_svabach_1 (CERtrain:0,29 %, CERvalidation: 2,33 %), bez existujúcich polygónov

Model 8: ID44251 MIX print_antikva_kurziva_fraktura_svabach_8 (počet slov: 1880 training set: pages 1 – 6, 9 – 10, validation set: pages 7 – 8) – vytvorený 24.8.2022, technológia HTR+, Nr. of Epochs 50, CER (train) = 0,17 %, s base modelom ID44001 MIX print_antikva_kurziva_fraktura_svabach_2 (CERtrain:0,25 %, CERvalidation: 3,09 %), bez existujúcich polygónov

a Bel_Adparatus_1735_model 3a boli zaradené 3 strany, WER1 a WER2 modelov preto v tabuľke uvádzame nasledovne: WER1 = najnižšia hodnota z troch, WER2 = najvyššia hodnota z troch.

Model 9a: ID45904 MIX print_antikva_kurziva_fraktura_svabach_9a (počet slov: 5818, training set: pages 1 – 7, 9, 11 – 20, validation set: pages 8, 10) – vytvorený 5.11.2022, pretrénovanie Modelu 8 technológiou ***PyLaia***, Nr. of Epochs 101, CER (train) = 1,00 %, bez base modelu, bez existujúcich polygónov

Model 9: ID46060 MIX print_antikva_kurziva_fraktura_svabach_9 (počet slov: 5791, training set: pages 1 – 7, 9, 11 – 20, validation set: pages 8, 10) – vytvorený 11.11.2022, vytrénovaný nanovo technológiou PyLaia, Nr. of Epochs 101, CER (train) = 2,80 %, bez base modelu, bez existujúcich polygónov, viac slov v cvičnom súbore v porovnaní s Modelom 8

Model 10: ID51087 MIX print_antikva_kurziva_fraktura_svabach_10 (počet slov: 5811, training set: pages 1 – 15, 17, 19 – 20, validation set: pages 16, 18) – vytvorený 1.4.2023, technológia PyLaia, Nr. of Epochs 153, CER (train) = 0,60 %, bez base modelu, bez existujúcich polygónov, bez použitia predvolenej funkcie *Deslant*

Model 11: ID60244 MIX print_antikva_kurziva_fraktura_svabach_11 (počet slov: 30892, training set: pages 1 – 5, 7, 9, 10, 12 – 19, 21, 22 – 32, 35 – 50, 52 – 60, 62 – 93, validation set: pages 6, 8, 11, 20, 23, 33, 34, 51, 61) – vytvorený 1.3.2024, technológia PyLaia, Nr. of Epochs 92, CERtrain = 0,60 %, bez base modelu, bez existujúcich polygónov (pôvodne s existujúcimi polygónmi, ale výsledok bol nepoužiteľný), trénovanie na všetkých stranách prepísaného a editovaného dokumentu

Bel_Adparatus_1735_model 1: ID385017 (počet slov: 10141, training set: pages 4 – 7, 9 – 12, 14 – 21, 23, 24, 42, 44, 46, 72, validation set: pages 8, 13, 22) – vytvorený 14.8.2025, technológia PyLaia, Nr. of Epochs 250, CER (train) = 0,32 %, s base modelom ID60244 Model 11 (CERtrain:0,60 %, CERvalidation: 1,00 %), s využitím existujúcich polygónov, s vylúčením slov označených ako unclear

Bel_Adparatus_1735_model 2: ID386277 (počet slov: 9215, training set: pages 4 – 7, 9 – 12, 14 – 21, 23, 24, 42, 44, 46, 72, validation set: pages 8, 13, 22) – vytvorený 16.8.2025, technológia PyLaia, Nr. of Epochs 250, CER (train) = 0,33 %, s base modelom ID60244 Model 11 (CERtrain:0,60 %, CERvalidation: 1,00 %), s využitím existujúcich polygónov, s vylúčením slov označených ako unclear, po oprave chýb v manuálnom prepise vzorky *Ground Truth*

Bel_Adparatus_1735_model 3: ID395837 (počet slov 9215, training set: pages: 4 – 7, 9 – 12, 14 – 21, 23, 24, 42, 44, 46, 72), validation set: pages: 8, 13, 22) – vytvorený 6.9.2025, technológia PyLaia, Nr. of Epochs 250, CER (train) = 0,31 %, s base modelom ID60244 Model 11 (CERtrain:0,60 %, CERvalidation: 1,00 %), bez existujúcich polygónov, s vylúčením slov označených ako unclear

Bel_Adparatus_1735_model 3a: ID405517 (počet slov 9215, training set: pages: 4 – 7, 9 – 12, 14 – 21, 23, 24, 42, 44, 46, 72), validation set: pages: 8, 13, 22) – vytvorený 22.9.2025, technológia PyLaia, Nr. of Epochs 250, CER (train) = 0,27 %, s base modelom ID60244 Model 11 (CERtrain:0,60 %, CERvalidation: 1,00 %), bez existujúcich polygónov, s vylúčením slov označených ako unclear; trénovaný s totožnými parametrami ako model 3 s časovým odstupom 15 dní na overenie (ne)stability AI