KOMENSKÝ A BEL V RUKÁCH UMELEJ INTELIGENCIE

Mgr. Michaela Mikušková; michaela.mikuskova@umb.sk; (Univerzitná knižnica Univerzity Mateja Bela v Banskej Bystrici)
Mgr. Lucia Nižníková; lucia.niznikova@umb.sk; (Univerzitná knižnica Univerzity Mateja Bela v Banskej Bystrici)

Účel – Článok sa zaoberá témou digital humanities a prináša jeden z príkladov dobrej praxe v tejto oblasti. Načrtáva hlavné ciele slovenského projektu SKRIPTOR, ktorý má národný aj európsky rozmer a priamo nadväzuje na excelentný európsky projekt READ. Jednoduchou formou predstavuje prácu so softvérom Transkribus Lite na rozpoznávanie a automatickú transkripciu historických dokumentov založenú na strojovom učení.

Metódy – Metodologický popis práce s platformou Transkribus Lite – nasnímanie dokumentu, segmentácia dokumentu, príprava vzorky Ground Truth, trénovanie modelu, vyhodnotenie modelu, automatická transkripcia celej zbierky. Na vyhodnotenie úspešnosti modelu sa použila miera chybovosti znakov (Character Error Rate – CER).

Výsledky – Sumarizuje výsledky dosiahnuté na špecifickej tlači Orbis Pictus (1798) a ich aplikáciu na podobnej tlači slovacikálnej proveniencie Adparatus Ad Historiam Hungariae z roku 1735 v prostredí platformy Transkribus. Stručne porovnáva výsledky transkripcie získané s využitím iného softvérového riešenia. Načrtáva postupy pre ďalšie spracovanie automaticky transkribovaného dokumentu.

Originalita/Hodnota – Projekt SKRIPTOR umožnil riešiteľom kompetentne participovať na výskumných aktivitách a svojimi výsledkami sa podieľať na udržateľnosti výsledkov výskumu európskeho projektu READ. Riešitelia už počas trvania projektu prispeli k tvorbe univerzálnych agregovaných modelov pre rukopisné a tlačené dokumenty v rámci Transkribus komunity.

http://doi.org/10.52036/1335793X.2024.1.15-27

Digitálne humanitné vedy (Digital Humanities, DH) sa zrodili po stretnutí tradičných humanitných vied s počítačovými metódami (Brudick et al. 2012). Táto oblasť vedeckej činnosti sa stále vyvíja a mení, striktné definície preto môžu rýchlo zastarať a prestať pokrývať všetko to, čo týmto pojmom na pomedzí digitálnych technológií a humanitných disciplín môžeme pomenovať. „Popri digitálnych archívoch, kvantitatívnych analýzach a projektoch vytvárania nástrojov, ktoré kedysi charakterizovali túto oblasť, DH v súčasnosti zahŕňajú širokú škálu metód a postupov: vizualizácie veľkých súborov obrázkov, 3D modelovanie historických artefaktov, ‚born digital' dizertačné práce, hashtagový aktivizmus a jeho analýzu, hry alternatívnej reality, mobilné makerspace a ďalšie. V takzvanom ‚veľkom stane' digitálnych humanitných vied môže byť niekedy ťažké konkrétne určiť, čo presne zahŕňa práca v tejto oblasti." (Klein 2016) Digitálne humanitné vedy vychádzajú z humanitnej informatiky (Humanities Computing), ktorej začiatky siahajú do minulého storočia. Digitálni humanisti ako prví osvojitelia technológií prezieravo vycítili, že výpočtová technika zohrá významnú úlohu vo výskume v humanitných vedách. Prvé vlny záujmu humanitných vied o počítačové siete a výpočtovú techniku dokladajú priekopnícke práce z konca 40. rokov a modely, ktoré inšpirovali archívne projekty v Oxforde na začiatku 70. rokov 20. storočia. V tom istom období „dajú sa zaznamenať určité rané i pokročilé etapy či stupne digital humanities v oblasti knihovníctva a informačných systémov, teda v oblasti knižničných a informačných systémov a služieb" aj na Slovensku (Katuščák 2020). V posledných rokoch sa disciplinárne zameranie DH rozšírilo o kritické digitálne štúdie a najmä o oblasti, ktoré sa častejšie spájajú s inžinierstvom, ako sú strojové učenie, dátová veda a umelá inteligencia (Berry 2019).

V oblasti digitálnych humanitných vied dnes existujú tisícky projektov od malých s obmedzeným alebo žiadnym financovaním až po rozsiahle projekty s viacročnou finančnou podporou. Niektoré pokračujú, iné skončia z dôvodu straty podpory alebo záujmu. Na projektoch digitálnych humanitných vied sa v porovnaní s tradičnými humanitnými a spoločenskými vedami zvyčajne podieľa väčší tím

riešiteľov, ktorý môžu tvoriť vedeckí a pedagogickí zamestnanci, študenti všetkých stupňov štúdia, odborníci na informačné technológie aj partneri z galérií, knižníc, archívov a múzeí. Autorstvo sa priznáva väčším skupinám ľudí a práve tento kolaboratívny charakter, ktorý možno prirovnať k prírodným vedám, odlišuje digital humanities od modelu výhradného autorstva typického pre tradičné humanitné vedy.

Za jedného z najvýznamnejších zástupcov projektov digitálnych humanitných vied (nielen) v Európe možno považovať projekt READ *The Recognition and Enrichment of Archival Documents*, ktorý v rokoch 2016 – 2019 získal podporu z programu Európskej únie Horizont 2020. V rámci tohto projektu konzorcium pod vedením profesora Güntera Mühlbergera z Univerzity v Innsbrucku vyvinulo technológiu rozpoznávania rukopisných textov Handwritten Text Recognition (HTR+) založenú na strojovom učení a vytvorilo platformu s názvom Transkribus na rozpoznávanie textu, jeho automatickú transkripciu a vyhľadávanie v historických dokumentoch. Krátko po skončení projektu sa Transkribus dostal na zoznam laureátov ceny Horizont Impact Award 2020 a pritiahol pozornosť odbornej i laickej verejnosti na celom svete. V súčasnosti sa ďalej rozvíja a využíva v rámci združenia READ-COOP SCE, ktoré má aktuálne 135 individuálnych a inštitucionálnych členov z 30 krajín. Jedným z nich je aj Univerzita Mateja Bela v Banskej Bystrici, ktorá v spolupráci so Štátnou vedeckou knižnicou v Banskej Bystrici v rámci projektu aplikovaného výskumu *Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov* (projekt SKRIPTOR) v období júl 2021 – jún 2024 overovala možnosti aplikácie nástroja Transkribus na rôzne typy historických rukopisných a tlačených textov slovacikálneho charakteru.

Hlavné ciele projektu možno zosumarizovať nasledovne:

- Oboznámiť sa s príkladmi dobrej praxe rozpoznávania historických dokumentov v Európe (najmä v krajinách ako Nemecko, Rakúsko, Španielsko, Fínsko, Dánsko, Veľká Británia a i.) a implementovať najnovšie poznatky z výskumu automatického rozpoznávania textov historických dokumentov na Slovensku. Podieľať sa na udržateľnosti výsledkov výskumu európskeho projektu READ a svojimi zbierkami a modelmi prispieť k tvorbe veľkých, tzv. super modelov.
- Prepojiť odborníkov z pamäťových inštitúcií, z oblasti informačných technológií a ďalších disciplín najmä humanitného a spoločenskovedného charakteru (história, jazykoveda, dokumentaristika) a ponúknuť riešenie, ktoré by mohlo byť základom pre funkčný národný e-archív ako nástroj na ochranu písomného dedičstva a jeho sprístupňovanie širšej odbornej i laickej verejnosti.
- Potvrdiť a dosiahnuť zníženie ceny transkripcie z 30 € za manuálnu transkripciu jednej strany na menej ako 1 €/strana za automatickú transkripciu textu. Automatický prepis s nástrojom Transkribus poskytuje v mnohých prípadoch texty s nízkou chybovosťou na úrovni diakritiky a interpunkcie, ktoré sú dobre čitateľné a zrozumiteľné. To sa dá dosiahnuť školením a trénovaním systému, ktorého výsledkom je vytvorenie modelu pre špecifický rukopis alebo zbierku. Je to zároveň jeden z najsilnejších argumentov na používanie platformy, pretože umožňuje každému používateľovi trénovať zodpovedajúce modely transkripcie presne podľa jeho požiadaviek.

Prudký vývoj digitálnych technológií a snaha rozširovať a vylepšovať možnosti a funkcionality systému vedú k rýchlym a zásadným zmenám na platforme Transkribus, ktoré sme pri prácach počas riešenia projektu Skriptor pociťovali s rôznou intenzitou. Z dôvodu časovej náročnosti na udržiavanie proprietárneho softvéru a zastaranosti kódu koncom roka 2022 pristúpil vývojový tím k odstaveniu technológie HTR+ a nahradil ju technológiou PyLaia. Aktuálne sú stále dostupné dve verzie *Transkribus Expert Client* (softvérová verzia, ktorá sa inštaluje priamo do osobného počítača) a *Transkribus Lite* (webová verzia, na ktorú je zameraný súčasný a budúci vývoj), avšak v lete 2023 vývojový tím avizoval ukončenie vývoja funkcií v prostredí expert klienta a presun vývoja na online verziu *Transkribus Lite/Beta*. Definitívne ukončenie verzie Transkribus Expert Client je naplánované na jún 2024. Hlavnou témou výročnej konferencie používateľov Transkribusu TUC 2024 boli veľké jazykové modely (*Large Language Models*), ktoré tvorcovia zatiaľ v testovacom režime skúšajú aktívne aplikovať do platformy.

V súčasnosti existuje množstvo nástrojov, ktoré využívajú metódu optického rozpoznávania znakov (Optical Character Recognition, OCR) a slúžia na prevod textu v obrázkoch do editovateľnej digitálnej formy dostupnej v textovom editore (BasuMallick 2023). Presnosť prepisu závisí od použitého softvéru. Samostatnou kapitolou sú softvéry schopné naučiť sa rozpoznávať rukou písaný text, jeho formu, štýl a špecifiká spojené so špeciálnymi grafémami. Sú schopné rozpoznávať diakritické znamienka, dobové pisárske značky, skratky aj sklon písma určitého rukopisu.

Softvér Transkribus je možné používať aj ako softvér pre OCR. Pre printové dokumenty má dostupné samostatné nástroje, ktoré uľahčujú rozpoznávanie textu s využitím tejto metódy (Using Transkribus for OCR with prin-

ted book 2023). Špecifické sú však jeho nástroje určené na rozpoznávanie ručne písaného textu (Handwritten Text Recognition, HTR), ktoré dokážu prepisovať rukopisné dokumenty (Dietrich 2021). Ako sme spomenuli v úvode, platformy na optické rozpoznávanie znakov a rozpoznávanie ručne písaného textu majú spoločnú jednu základnú črtu: obe konvertujú obraz dokumentu na text. Na platformu sa nahrá naskenovaná stránka z knihy a tá prekonvertuje vytlačené slová na digitálny textový súbor. Prednosťou Transkribusu v porovnaní s inými softvérmi je predovšetkým možnosť vytvárania vlastných modelov na automatickú transkripciu rôznych typov dokumentov (rukopisných, strojopisných aj tlačených).

Riešiteľský kolektív projektu SKRIPTOR pracoval s rukopisnými, tlačenými a strojopisnými zbierkami. Členky riešiteľského kolektívu Univerzitnej knižnice Univerzity Mateja Bela v Banskej Bystrici upriamili pozornosť na dve historické tlače z vlastného fondu knižnice: *Orbis Pictus* Jána Amosa Komenského z roku 1798 vydaný v kníhtlačiarni Šimona Petra Webera v Bratislave a *Adparatus ad Historiam Hungariae* Mateja Bela vydaný v roku 1735 u Jána Pavla Royera v Bratislave.

Orbis Pictus je 184-stranová publikácia napísaná v štyroch jazykoch, pričom každý jazyk je vytlačený iným fontom – latinský antikvou, maďarský kurzívou, nemecký fraktúrou a český švabachom. Text je formálne usporiadaný do štyroch blokov. Latinská, maďarská a nemecká mutácia je zoradená v stĺpcoch a česká má klasickú knižnú podobu. Súčasťou každej kapitoly je ilustrácia v podobe drevorytu, ku ktorej sa následne viaže textový popis. Každý jazyk obsahuje špecifické grafémy, ktoré sa dnes už nepoužívajú. V latinčine ide najmä o ligatúry æ, œ, v maďarčine o znaky ö, ű, v nemčine o ligatúry ß, ʒ a znaky å, o̊, ů. Najviac špecifických znakov je obsiahnutých v češtine ë, š, ċ, đ, ė, ġ, ṅ, ř, ṡ, ṫ, ż. Vo všetkých štyroch jazykoch sa používa graféma dlhé s (ſ). Z hľadiska skúmania možností softvéru Transkribus a tvorby modelu automatickej transkripcie bola výzvou práve jazyková a typografická rôznorodosť dokumentu a spôsob usporiadania textu.

Adparatus ad Historiam Hungariae má 728 strán. Napísaný je v latinskom jazyku s použitím dvoch fontov – antikvy a kurzívy. Text dopĺňajú marginálie, niektoré časti majú stĺpcové usporiadanie, napr. poznámky k textu. Kapitoly sa začínajú ozdobnou iniciálou. V texte sa nachádzajú špecifické ligatúry, napr. ct a väčšie množstvo textu je napísaného kapitálkami. Tento dokument bol zaujímavý z hľadiska overenia funkčnosti vlastného vytrénovaného modelu automatickej transkripcie a jeho prípadného vylepšenia.

PRÍPRAVA DIGITALIZÁTU

Prvým krokom v procese práce so softvérom Transkribus je príprava kvalitných snímok dokumentu. Z kvalitatívneho hľadiska sú najlepšie digitalizáty vyhotovené na skeneri. Softvér však dokáže spoľahlivo fungovať aj na snímkach zhotovených pomocou smartfónu. Odporúčaná kvalita skenovania je minimálne 600 DPI, pri rukopisoch 900 – 1200 DPI. Softvér pracuje s dokumentmi vo formáte JPG, PNG a PDF.

Vývojový tím Transkribusu vytvoril dva inovatívne nástroje, ktoré umožňujú rýchle a cenovo dostupné snímanie dokumentov:

- **ScanTent** – prenosný plátenný stan. Vďaka svojmu tvaru, použitým materiálom a integrovanému LED osvetleniu poskytuje optimálne svetelné podmien-

Obr. 1 ScanTent. Zdroj: help.transkribus.com

ky na snímanie dokumentu pomocou smartfónu až do veľkosti formátu A3 (obr. 1).

- **DocScan** – aplikácia určená na digitalizáciu dokumentov pomocou smartfónov. Možno ju použiť na akýkoľvek druh dokumentu, obzvlášť užitočná je na knihy, pretože automaticky zaregistruje otočenie strany a vytvorí snímku bez nutnosti opakovaného stláčania tlačidla na snímanie. Zároveň slúži na spracovanie nasnímaných obrázkov (orezanie, otočenie) a ich nahrávanie na server Transkribusu.

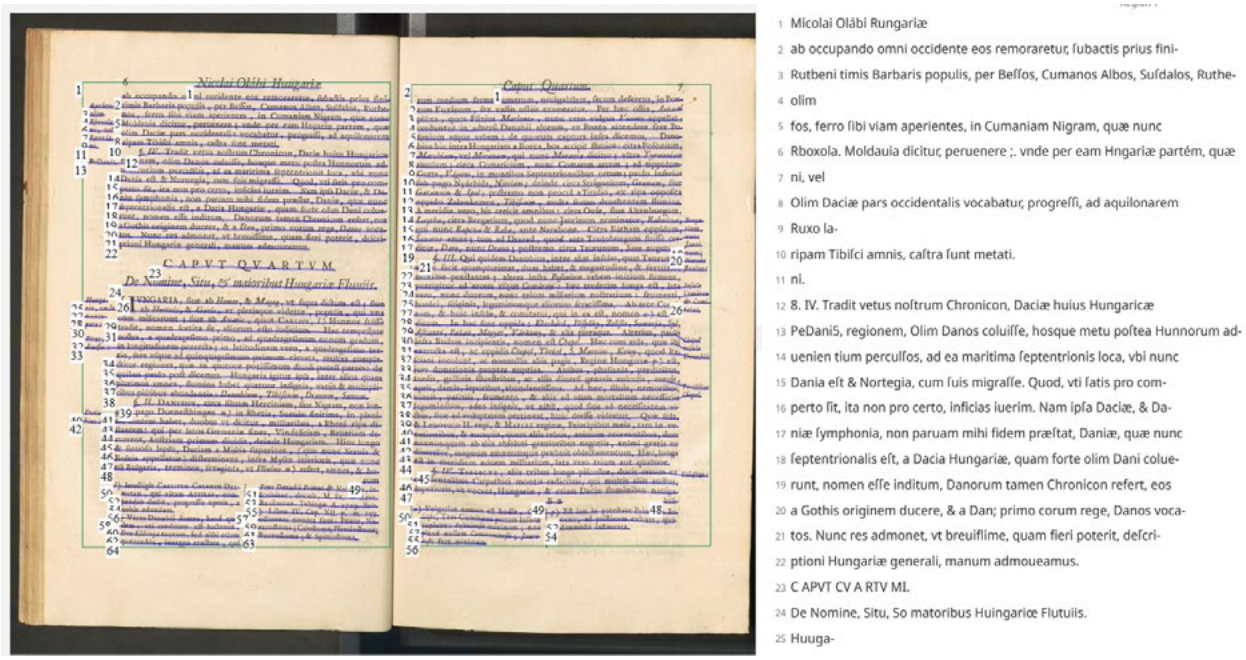
Na ďalšiu prácu s dokumentom a jeho prepis treba zbierku nahrať do vlastného používateľského účtu na platforme *Transkribus Lite* (https://readcoop.eu/transkribus). Registrácia je bezplatná. Zbierky sa ukladajú na serveroch spoločnosti READ-COOP SCE s prednastavenými prístupovými právami na súkromné použitie, t. j. nie sú viditeľné pre ostatných používateľov. Každý používateľ môže manažovať a povoľovať prístup k svojej zbierke iným používateľom podľa vlastného uváženia.

SEGMENTÁCIA DOKUMENTU

Nástroje slúžiace na rozpoznávanie textu na snímkach a jeho následný automatický prepis pracujú v dvoch fázach. Prvá je rozpoznanie stránky alebo analýza jej rozloženia (*Layout Analysis*), druhá je manuálny prepis textu v identifikovanej štruktúre snímky. Výsledkom analýzy rozloženia je segmentácia snímky dokumentu, ktorou sa identifikujú:

- jednotlivé objekty – textové rámce (*Text Regions*), riadky (*Lines*), obrázky, grafické prvky a i.;
- štruktúra rozloženia objektov na snímke;
- spôsob orientácie textu – horizontálny, vertikálny, zmiešaný;
- poradie čítania textu – zvyčajne z ľavého horného rohu smerom k pravému dolnému rohu.

Softvér Transkribus ponúka nástroje na segmentáciu textových rámcov, riadkov a tabuliek. Úplnú alebo čiastočnú segmentáciu je možné nechať výlučne na softvér. Dostupné sú hotové modely na segmentáciu určitých typov dokumentov (noviny, pohľadnice a i.), dokonca sa dá vytrénovať vlastný model pre špecifický typ dokumentu (Baseline model [2023]). V jednom kroku je možné vykonať segmentáciu textu a jeho automatický prepis (obr. 2). Oba dokumenty, s ktorými sme pracovali, mali špecifickú štruktúru. *Orbis Pictus* obsahoval text v stĺpcovom rozložení, ktorý je možné segmentovať aj pomocou nástrojov na segmentáciu tabuliek. *Adparatus* obsahoval marginálie, ktoré bolo treba oddeliť od hlavného textu a následne označiť poradie čítania jednotlivých rámcov podľa toho, ktorej časti textu sa marginálie týkajú. Na niektorých stranách boli aj poznámky k textu, ktoré sa nachádzali pod hlavným textom v dvoch stĺpcoch. Keďže automatická segmentácia textových rámcov nebola uspokojivá, osvedčilo sa nám manuálne označenie textových rámcov na jednotlivých snímkach a následné spustenie automatickej segmentácie riadkov v nich.



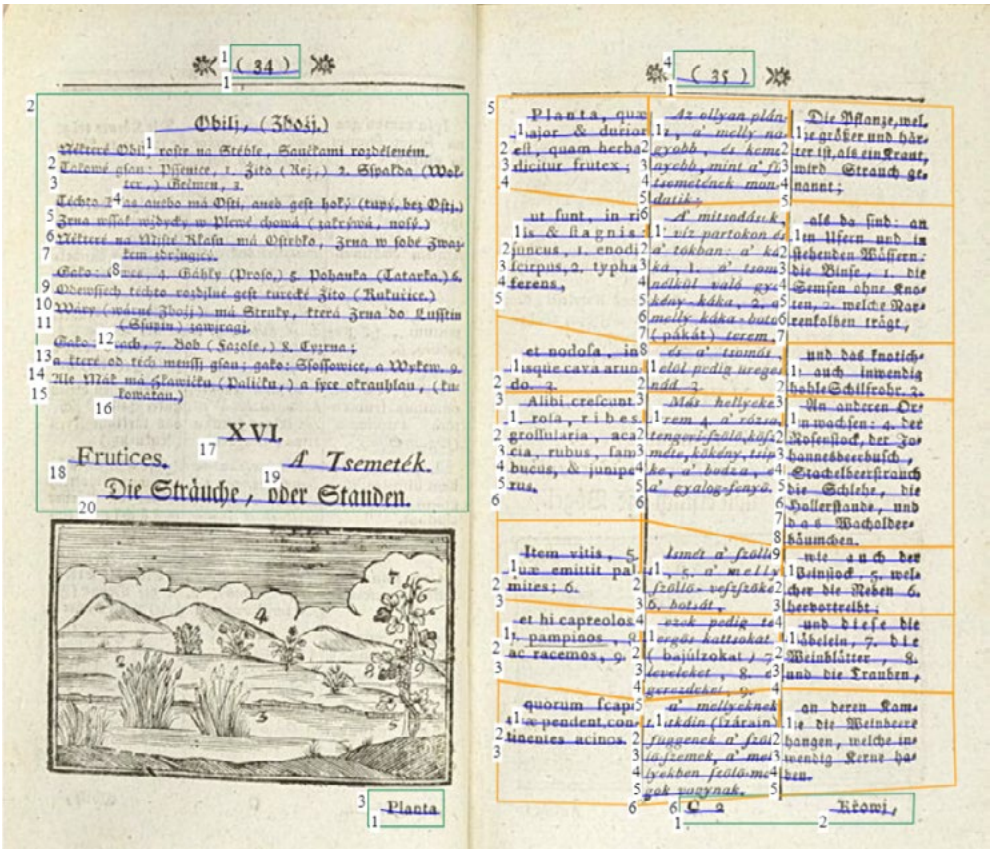
Obr. 2 Ukážka automatickej segmentácie a prepisu dokumentu s využitím nami vytrénovaného modelu na dokumente Adparatus. Vzhľadom na štruktúru dokumentu nie sú správne segmentované textové rámce, chybné je aj poradie čítania jednotlivých riadkov. Zdroj: Transkribus

Transkribus ponúka užitočné nástroje na konfiguráciu automatickej segmentácie prostredníctvom rozšírených nastavení. Na segmentáciu riadkov v už zadefinovaných textových rámcoch treba v nastavení Generovanie textových rámcov (*Generation of Text Regions*) vybrať možnosť Ponechať existujúce textové rámce (*Keep existing*). Pri tlačiach, ktorých konce riadkov sa nachádzajú v tesnej blízkosti hraníc dvoch susediacich textových rámcov, môže v procese automatickej segmentácie dochádzať k spájaniu riadkov medzi textovými rámcami. Pri práci s našou tlačou sa preto osvedčilo použitie funkcie Rozdeliť riadky v textových rámcoch (*Split lines on regions*), ktorá zabraňuje prepájaniu riadkov medzi dvomi textovými rámcami.

MANUÁLNY PREPIS VYBRANÝCH STRÁN

Po segmentácii dokumentu nasleduje manuálny prepis vybraných strán. Slúži ako predloha, podľa ktorej sa stroj „učí" čítať nasegmentovaný text a alfanumerické znaky, z ktorých pozostáva. Predpokladom vytvorenia modelu na transkripciu je príprava vzorky *Ground Truth* (tzv. základná pravda). V strojovom učení predstavuje presné a objektívne informácie, ktoré sa považujú za správne. Tvorí základ na trénovanie modelu a následné vyhodnotenie jeho presnosti porovnaním získaných výsledkov so vzorkou *Ground Truth*. Je dôležité, aby strany vo vzorke boli dostatočne reprezentatívne a obsahovali potrebné množstvo všetkých typov písma, ktoré sú v dokumente zastúpené. *Ground Truth* má významný vplyv na kvalitu vytrénovaného modelu.

V procese práce s dokumentom softvér Transkribus ponúka niekoľko možností označenia stavu spracovania dokumentu. Stav editácie konkrétnej stránky je možné označovať verziami dostupnými v hornej lište pracovného okna. Každá snímka nahratá do Transkribusu má označenie nová (*New)*. Po prvej editácii snímky, napr. po manuálnej segmentácii, sa snímka automaticky ukladá v stave prebiehajúci (*In Progress*). Ďalšie stavy si manažuje používateľ podľa toho, v akom stave prepisu sa daná strana nachádza. Stav hotový (*Done*) sa používa na označenie prepisu, ktorý neprešiel kontrolou, a stav finálna verzia (*Final*) na označenie prepísanej a skontrolovanej strany dokumentu. Trénovanie modelov sa vykonáva na snímkach označených ako *Ground Truth*. Do tohto stavu sa strana ukladá vtedy, keď sa jej prepis



Obr. 3 Ukážka segmentácie na snímke z tlače Orbis Pictus. Textové rámce boli segmentované manuálne, text v stĺpcoch pomocou nástroja Tabuľka (oranžovou farbou). Segmentáciu riadkov vykonal softvér automaticky na základe nastavených parametrov. Zdroj: Transkribus

už nebude meniť a môže sa použiť na trénovanie modelu a vyhodnotenie jeho kvality.

Vo fáze manuálneho prepisu vzorky treba dodržať nasledujúce zásady:

- Výber strán má byť čo najuniverzálnejší tak, aby reprezentatívne obsahoval všetky typy písiem v dokumente (fonty v tlačených dokumentoch, všetky „ruky" v rukopisných dokumentoch) vrátane jazykov, abecedných sád a štýlov písania (aj rukopis jedného autora sa môže počas života meniť).
- Snímky by nemali mať machule, nečistoty a ideálne ani presvit textu z druhej strany. Nečitateľné miesta sa dajú označiť tagom *Unclear*, čím sa vylúčia z procesu trénovania a negatívne neovplyvnia úspešnosť modelu.
- V závislosti od zamýšľanej miery zachovania autenticity originálu možno na prepis použiť metódu *transkripcie* (zjednodušene to znamená nahradenie starších a nepoužívaných znakov modernými) alebo metódu *transliterácie*. Transliterácia alebo odborný/vedecký prepis má niekoľko definícií, v zásade však platí, že ide o „prevod z jednej grafickej sústavy do druhej, pri ktorom každému písmenu jedného grafického systému zodpovedá vždy písmeno druhého systému (rovnaké písmeno alebo spojenie písmen), takže je možný aj jednoduchý spätný prevod do jazyka originálu" (Mistrík 1993, s. 343). Manuály na prácu so softvérom Transkribus odporúčajú prepis systémom písmeno za písmeno a v prípade stredovekých textov nepoužívanie modernej interpunkcie, v takýchto prípadoch je vhodnejšie interpunkciu úplne vynechať alebo používať špeciálne symboly (podrobne o pravidlách transkripcie v prostredí Transkribusu pozri Keijser a Moss 2019). Keďže transliterácia (paleografický prepis) sa primárne orientuje na vizuálnu podobu pôvodného textu (hoci často zachytáva aj fonetickú hodnotu pôvodnej grafémy) a zároveň poskytuje informáciu o tom, aké grafémy obsahuje pôvodný text, umožňuje rekonštrukciu písanej podoby slova. Výsledný prepis metódou transliterácie tak môže byť atraktívny nielen pre historikov a literárnych vedcov, ale aj pre jazykovedcov a grafológov. Zásady diakritiky podrobnejšie popisuje príručka *Automatická transkripcia historických dokumentov* (Katuščák a Nagy 2023, s. 85 – 90).
- Manuálny prepis vzorky, ktorý sa realizuje v textovom editore v pravej polovici obrazovky, musí byť čo najpresnejší a mal by kopírovať originálny text so všetkými chybami a omylmi (úlohou prepisu nie je korektúra pôvodného textu). Akékoľvek odchýlenie od originálu sa môže prejaviť chybovosťou v trénovanom modeli. Každá definitívne prepísaná strana sa označí príznakom *Ground Truth*.

Na prepis špeciálnych grafém sú k dispozícii bežné sady UNICODE, ktoré sa pri prepise vkladajú do textu pomocou virtuálnej klávesnice (nástroj *Virtual Keyboard*). Ak v ponuke klávesnice nejaké znaky chýbajú, dajú sa na platformu doplniť. Na generovanie kombinovaných UNICODE znakov odporúčame využiť niektorý z nástrojov Online Unicode Tools dostupných na internete.

TRÉNOVANIE MODELU

Počet strán s označením *Ground Truth* potrebných na trénovanie závisí od typu dokumentu. Pri tlačených dokumentoch postačuje rozsah okolo 5 000 slov, približne 20 – 30 strán. V prípade rukopisných dokumentov by mala vzorka obsahovať minimálne 15 000 slov, odhadom 50 – 75 strán (Data Preparation [2023]).

Takto pripravená tréningová vzorka (*Training Data*) sa v odporúčanom pomere 10:1 rozdelí pomocou funkcie *Train New Model* na cvičný súbor (*Training Set*) a overovací súbor (*Validation Set*). Na cvičnom súbore sa stroj učí, pri každom cykle prečíta rovnaké strany a v nasledujúcom cykle chybne rozpoznané znaky vyradí. Tieto údaje o správnom a nesprávnom čítaní sa stávajú základom modelu. Po vytrénovaní Transkribus na stranách zaradených do cvičného súboru model automaticky preverí na overovacom súbore, ktorý slúži na praktické odskúšanie modelu.

Vstupné údaje, ktoré používateľ zadáva pred spustením modelu, obsahujú názov modelu, jeho popis, jazyk dokumentu ako povinný údaj a storočie vydania dokumentu ako odporúčaný údaj. V tejto fáze je možné do trénovania pribrať základný model (*Base Model*). Pridanie dát tohto modelu umožní trénovať s menšou vzorkou *Ground Truth* a za určitých podmienok zlepšiť trénovaný model. Ako base model je možné použiť predchádzajúci vlastný model alebo verejne dostupný model iného používateľa trénovaný na dokumente s podobnými charakteristikami. Keďže hlavnou snahou komunity okolo nástroja Transkribus je tvorba univerzálnych agregovaných modelov, neustále sa zvyšujúci počet používateľov a pribúdanie dokumentov v zbierkach umožnil v posledných mesiacoch vytvorenie prvých veľkých modelov, tzv. super modelov ako *The Text Titan I, Dutch Demeter I, Amsterdam Notarian, Notarial 17th Century Super Model*, ktoré sú trénované na miliónoch rukopisných strán, a ďalšie veľké modely ako pre rukopisné (napr. *The Dutchess, The German Giant I, Transkribus French*

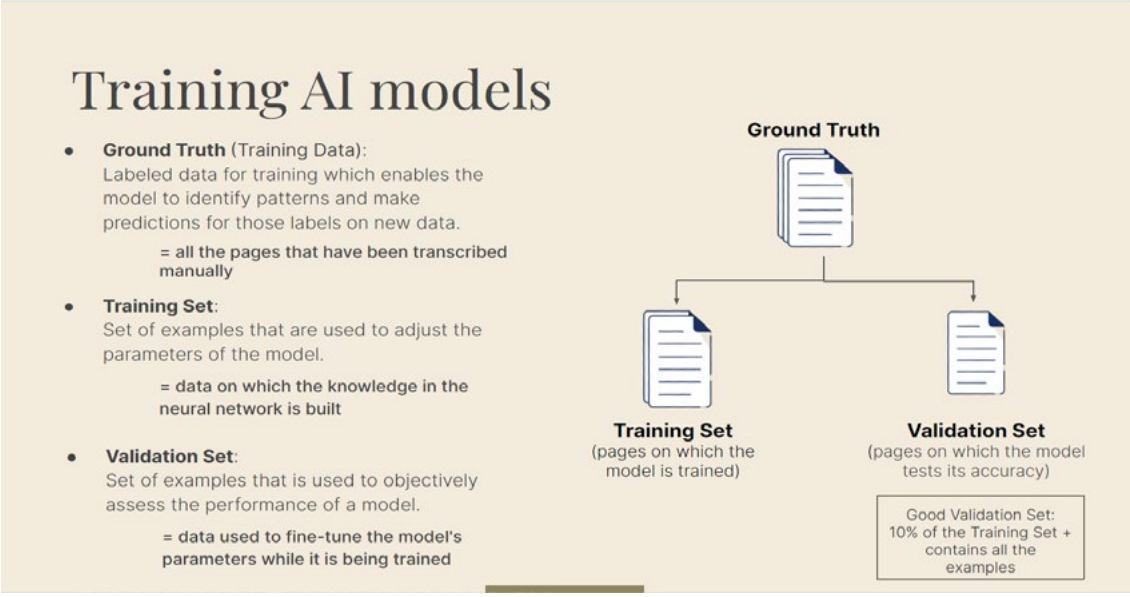
Model a i.), tak aj pre tlačené dokumenty (viac informácií na https://readcoop.eu/transkribus/public-models/). Aktuálne najväčší agregovaný model pre tlače *Transkribus Print M1* bol trénovaný na viac ako piatich miliónoch strán s chybovosťou len 2,20 %. Podľa dostupných údajov k modelu boli do trénovania zahrnuté aj dokumenty v slovenskom a českom jazyku, a keďže autorstvo modelu je pripísané Transkribus komunite, je predpoklad, že na jeho vytrénovanie boli použité aj segmentované strany tlačených dokumentov, ktoré sme zdigitalizovali, importovali na platformu a následne s nimi pracovali. Tým sme naplnili jeden z cieľov projektu, ktorým je prispenie k zvýšeniu efektívnosti nástroja Transkribus.

Prostredníctvom rozšírených nastavení (*Advanced Settings*) sa dá meniť maximálny počet cyklov, v ktorých sa má trénovanie (učenie) opakovať (*Training Cycles*), nastaviť minimálny počet opakovaní cyklov (*Early Stopping*) či vynechať z trénovania nečitateľné alebo nejasné miesta v texte.

VYHODNOTENIE MODELU

Po skončení trénovania Transkribus ponúkne vyhodnotenie v podobe percentuálneho vyjadrenia chybovosti a grafu. Vyhodnotenie obsahuje názov modelu, jeho jedinečné identifikačné číslo na platforme (*Model ID*), počet slov v cvičnom súbore (*Training Set Size*) a mieru chybovosti znakov CER (*Character Error Rate*) na overovacom súbore

Obr. 4 Ukážka prepisu originálu v textovom editore. Zdroj: Transkribus

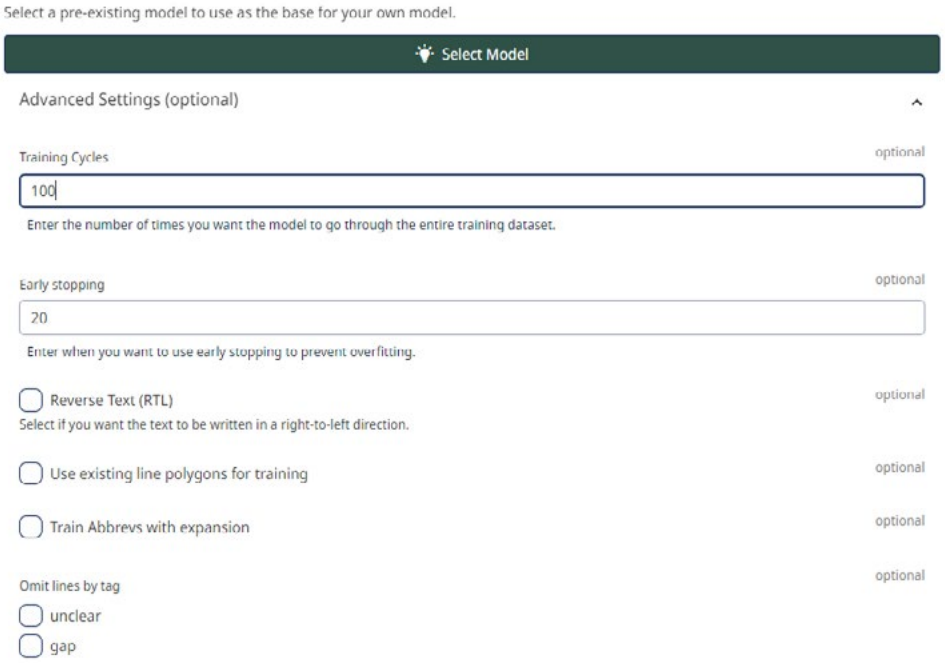


Obr. 5 Rozdelenie vzorky na trénovanie modelu. Zdroj: Transkribus

vyjadrenú v percentách. 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*. Vypočítava sa podľa vzorca CER = [(i + s + d) / n]*100. Každá malá chyba pri prepise je štatisticky plnohodnotnou chybou. To znamená, že chýbajúca čiarka, *u* namiesto *v*, *á* namiesto *ä*, medzera navyše alebo veľké písmeno namiesto malého sa počítajú ako chyby. Na základe doterajších skúseností Transkribus komunity sa modely s chybovosťou do 5 % pri tlačených textoch a do 10 % pri rukopisných dokumentoch považujú za vynikajúce.

lu aj na úrovni jednotlivých strán pomocou porovnania textových verzií (aktuálne je táto funkcionalita dostupná len vo verzii Expert klient). Používateľ získa podrobný prehľad o tom, čo model prepísal správne a kde v porovnaní s verziou *Ground Truth* urobil chybu. Z porovnania verzií na obr. 8 je vidieť, že prepísaný text obsahuje iba chyby v interpunkcii a diakritike, inak je dobre čitateľný a zrozumiteľný (Character Error Rate and Learning Curve [2023]).

Model 11 má zo všetkých modelov, ktoré sme trénovali na tlači J. A. Komenského *Orbis Pictus* (1798), najlepšie parametre, v zmysle kritérií vyhodnocovania úspešnosti ho teda možno považovať za vynikajúci (podrobne o tré-



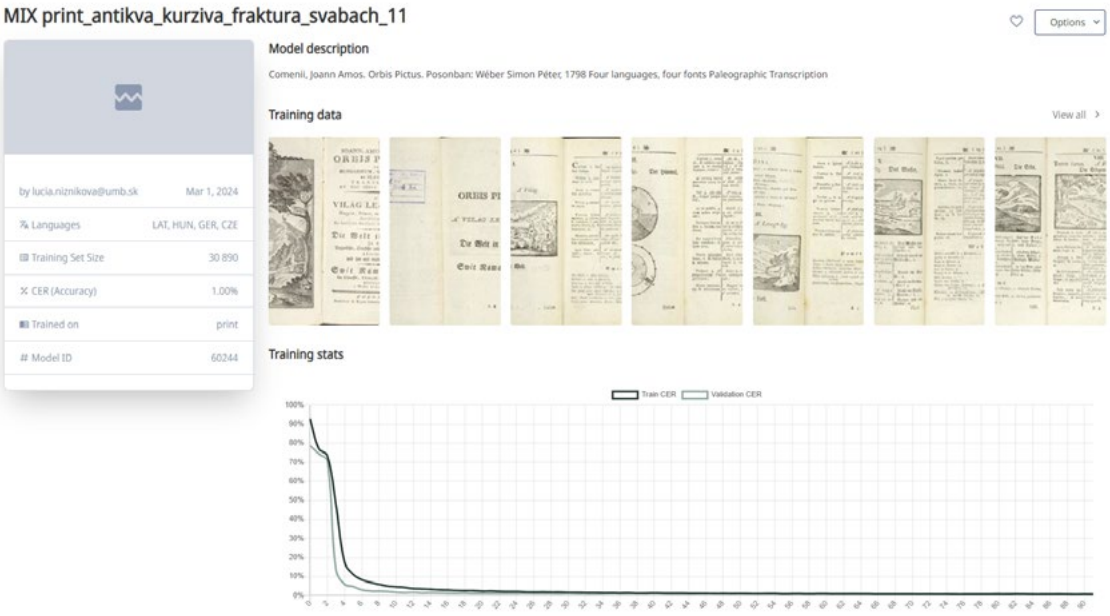
Obr. 6 Nastavenie parametrov na trénovanie modelu

V grafe (*Training Stats*) sú zobrazené dve čiary. Čierna čiara predstavuje priebeh trénovania, zelená čiara predstavuje priebeh vyhodnocovania naučeného na overovacom súbore. Os y zobrazuje mieru chybovosti znakov CER (*Character Error Rate*). Krivka sa vždy začína na 100 % a ako sa model trénuje a zlepšuje, postupne klesá. Os x predstavuje cykly, t. j. priebeh tréningu. Počas procesu trénovania vykonáva Transkribus vyhodnotenie po každom cykle. Model na obrázku bol vytrénovaný pri počte 92 cyklov. V tomto prípade bol maximálny počet cyklov nastavený na 250, ale trénovanie sa automaticky zastavilo skôr, pretože model sa už nezlepšoval.

Hodnota CER 1,00 % na obr. 7 znamená, že v cvičnom súbore bolo bezchybne rozpoznaných až 99 % alfanumerických znakov. Transkribus ponúka vyhodnotenie modelu aj na úrovni jednotlivých strán pomocou porovnania textových verzií (aktuálne je táto funkcionalita dostupná len vo verzii Expert klient). Používateľ získa podrobný prehľad o tom, čo model prepísal správne a kde v porovnaní s verziou *Ground Truth* urobil chybu. Z porovnania verzií na obr. 8 je vidieť, že prepísaný text obsahuje iba chyby v interpunkcii a diakritike, inak je dobre čitateľný a zrozumiteľný (Character Error Rate and Learning Curve [2023]).

Model 11 má zo všetkých modelov, ktoré sme trénovali na tlači J. A. Komenského *Orbis Pictus* (1798), najlepšie parametre, v zmysle kritérií vyhodnocovania úspešnosti ho teda možno považovať za vynikajúci (podrobne o trénovaní a výsledkoch predchádzajúcich modelov Nižníková a Mikušková 2022). Jediný spôsob, ako zistiť, do akej miery je aj univerzálny a ďalej využiteľný, je odskúšať ho na automatickej transkripcii iného dokumentu s podobnými vlastnosťami (historické obdobie, jazyk, typ písma/font a i.). Na tento účel sme vybrali tlačené dielo Mateja Bela *Adparatus ad Historiam Hungariae* z roku 1735, pretože zodpovedá uvedeným kritériám: pochádza z toho istého storočia, je písané po latinsky a pri tlači boli použité fonty antikva a kurzíva.

Výsledky automatickej transkripcie vybraných strán *Adparatu* modelom 11 sú uspokojivé. Keďže však model 11 nebol trénovaný na dostatočnej vzorke textu písaného kapitálkami, práve tam sme zaznamenali vysokú chybovosť. Úvodné strany dokumentu sú vytlačené iným ty-



Obr. 7 Vyhodnotenie úspešnosti modelu 11 (MIX print_antikva_kurziva_fraktura_svabach_11)

pom antikvy ako hlavné časti textu. Na prvý pohľad je na týchto stranách zreteľný aj rozdiel v kvalite papiera a výrazný presvit textu z druhej strany listu. Všetky tieto faktory mali pravdepodobne veľký vplyv na vyššiu chybovosť prepisu za použitia vytrénovaného modelu.

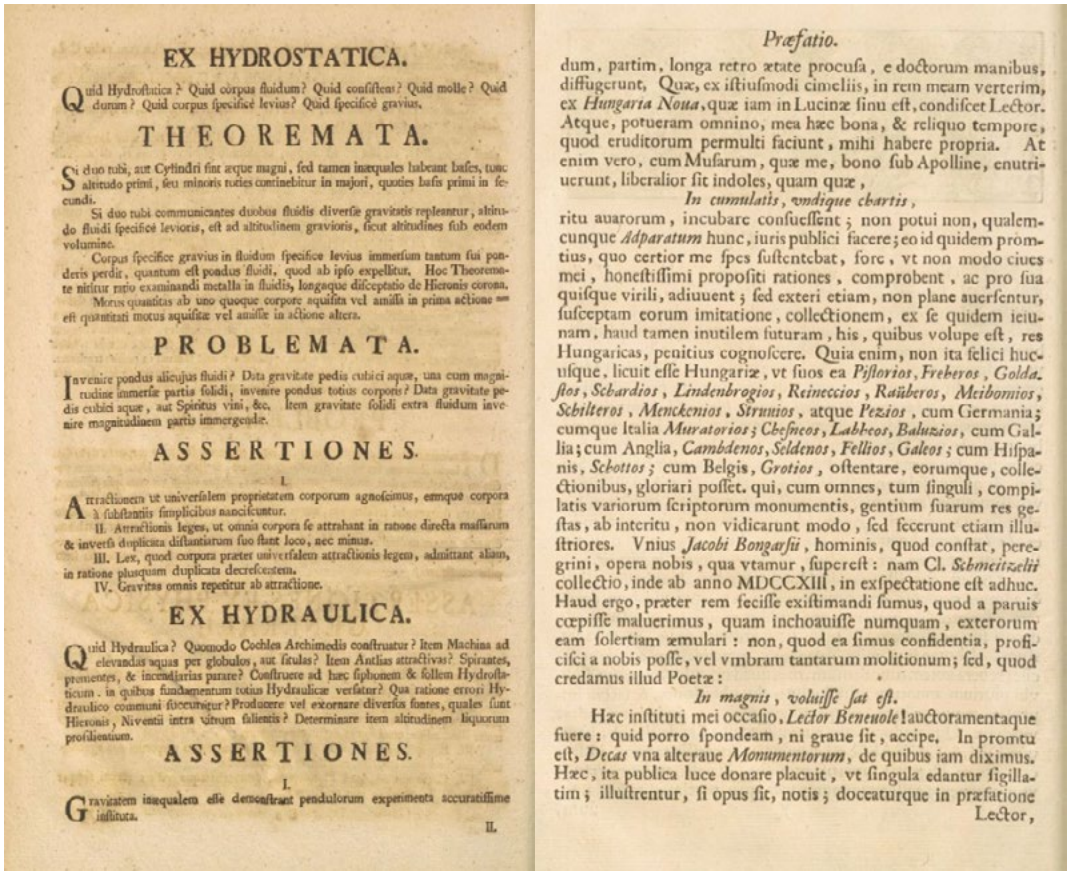
MOŽNOSTI PRÁCE S TRANSKRIBOVANÝM TEXTOM

Po automatickej transkripcii dokumentu je možné s prepísaným textom ďalej pracovať. Dokument možno obohatiť o dodatočné informácie (metadáta). Prehľadnosť dôležitých údajov nachádzajúcich sa v texte a orientáciu v štruktúre dokumentu zjednodušuje použitie tagov. Transkribus ponúka dva typy tagov:

- *textové tagy* definujú pojmy a frázy v texte, napr. osobné mená, geografické označenia, dátumy, skratky a i.,
- *štrukturálne tagy* definujú štruktúru dokumentu a vzťahujú sa na oblasti textu a riadkov, napr. nadpis, odsek, čísla strán, kustódy, marginálie a i.

Zbierky a ich prepisy sa dajú z platformy Transkribus Lite vyexportovať v niektorom z ponúkaných formátov (PDF,

Obr. 8 Ukážka porovnania automaticky prepísaného textu (chyby zvýraznené červenou farbou) s predlohou Ground Truth (zelenou farbou). Zdroj: Transkribus



EX HYDROSTATICA.

Quid Hydroſtatica? Quid corpus fluidum? Quid conſiſtens? Quid molle? Quid durum? Quid corpus ſpecifice levius? Quid ſpecifice gravius.

THEOREMATA.

Si duo tubi, aut Cylindri ſint æque magni, ſed tamen inæquales habeant baſes, tunc altitudo primi, ſeu minoris toties continebitur in majori, quoties baſis primi in ſecundi.

Si duo tubi communicantes duobus fluidis diverſæ gravitatis repleantur, altitudo fluidi ſpecificè levioris, eſt ad altitudinem gravioris, ſicut altitudines ſub eodem volumine.

Corpus ſpecifice gravius in fluidum ſpecifice levius immerſum tantum ſui ponderis perdit, quantum eſt pondus fluidi, quod ab ipſo expellitur. Hoc Theoremate nititur ratio examinandi metalla in fluidis, longaque diſceptatio de Hieronis corona.

Motus quantitas ab uno quoque corpore aquiſita vel amiſſa in prima actione = eſt quantitati motus aquiſitæ vel amiſſæ in actione altera.

PROBLEMATA.

Invenire pondus alicujus fluidi? Data gravitate pedis cubici aquæ, una cum magnitudine immerſæ partis ſolidi, invenire pondus totius corporis? Data gravitate pedis cubici aquæ, aut Spiritus vini, &c. Item gravitate ſolidi extra fluidum invenire magnitudinem partis immergendæ.

ASSERTIONES.

I.

Attractionem ut univerſalem proprietatem corporum agnoſcimus, eamque corpora à ſubſtantiis ſimplicibus nanciſcuntur.

II. Attractionis leges, ut omnia corpora ſe attrahant in ratione directa maſſarum & inverſa duplicata diſtantiarum ſuo ſtant loco, nec minus.

III. Lex, quod corpora præter univerſalem attractionis legem, admittant aliam, in ratione pluſquam duplicata decreſcentem.

IV. Gravitas omnis repetitur ab attractione.

EX HYDRAULICA.

Quid Hydraulica? Quomodo Cochlea Archimedis conſtruatur? Item Machina ad elevandas aquas per globulos, aut ſiculas? Item Antlias attractivas? Spiraates, prunentes, & incendiarias parare? Conſtruere ad hæc ſiphonem & ſollem Hydroſtaticum, in quibus fundamentum totius Hydraulicæ verſatur? Qua ratione errori Hydraulico communi ſuccurritur? Producere vel exornare diverſos fontes, quales ſunt Hieronis, Niventii intra vitrum ſalientis? Determinare item altitudinem liquorum proſilientium.

ASSERTIONES.

I.

Gravitatem inæqualem eſſe demonſtrant pendulorum experimenta accuratiſſime inſtituta.

II.

Præfatio.

dum, partim, longa retro ætate procuſa, e doctorum manibus, diffugerunt. Quæ, ex iſtiuſmodi cimeliis, in rem meam verterim, ex *Hungaria Noua*, quæ iam in Lucinæ ſinu eſt, condiſcet Lector. Atque, potueram omnino, mea hæc bona, & reliquo tempore, quod eruditorum permulti faciunt, mihi habere propria. At enim vero, cum Muſarum, quæ me, bono ſub Apolline, enutrierunt, liberalior ſit indoles, quam quæ,

In cumulatis, vndique chartis,

ritu auarorum, incubare conſueſſent; non potui non, qualemcunque *Adparatum* hunc, iuris publici facere; eo id quidem promtius, quo certior me ſpes ſuſtentebat, fore, vt non modo ciues mei, honeſtiſſimi propoſiti rationes, comprobent, ac pro ſua quiſque virili, adiuuent; ſed exteri etiam, non plane auerſentur, ſuſceptam eorum imitatione, collectionem, ex ſe quidem ieiunam, haud tamen inutilem futuram, his, quibus volupe eſt, res Hungaricas, penitius cognoſcere. Quia enim, non ita felici hucuſque, licuit eſſe Hungariæ, vt ſuos ea *Piſtorios, Freberos, Goldaſtos, Schardios, Lindenbrogios, Reineccios, Raüberos, Meibomios, Schilteros, Menckenios, Struuios*, atque *Pezios*, cum Germania; eumque Italia *Muratorios; Cheſneos, Labbeos, Baluzios*, cum Gallia; cum Anglia, *Cambdenos, Seldenos, Fellios, Galeos*; cum Hiſpanis, *Schottos*; cum Belgis, *Grotios*, oſtentare, eorumque, collectionibus, gloriari poſſet. qui, cum omnes, tum ſinguli, compilatis variorum ſcriptorum monumentis, gentium ſuarum res geſtas, ab interitu, non vindicarunt modo, ſed fecerunt etiam illuſtriores. Vnius *Jacobi Bongarſii*, hominis, quod conſtat, peregrini, opera nobis, qua vtamur, ſupereſt: nam Cl. *Schmeitzelii* collectio, inde ab anno MDCCXIII, in exſpectatione eſt adhuc. Haud ergo, præter rem feciſſe exiſtimandi ſumus, quod a paruis cœpiſſe maluerimus, quam inchoauiſſe numquam, exterorum eam ſolertiam æmulari: non, quod ea ſimus confidentia, proficiſci a nobis poſſe, vel vmbram tantarum molitionum; ſed, quod credamus illud Poetæ:

In magnis, voluiſſe ſat eſt.

Hæc inſtituti mei occaſio, *Lector Beneuole* auctoramentaque fuere: quid porro ſpondeam, ni graue ſit, accipe. In promtu eſt, *Decas* vna alteraue *Monumentorum*, de quibus iam diximus. Hæc, ita publica luce donare placuit, vt ſingula edantur ſigillatim; illuſtrentur, ſi opus ſit, notis; doceaturque in præfatione

Lector,

Obr. 9 Ukážka rozdielnej kvality papiera a tlače na stránkach Adparatu

TXT, docx, XML, obrázok) a pracovať s nimi mimo platformy. Exportovať možno aj metadáta v podobe textových a štrukturálnych tagov.

Na zverejnenie originálnych dokumentov doplnených prepísanou verziou v online prostredí softvér Transkribus ponúka špecializovanú platformu *Transkribus Sites* (2024). Je navrhnutá tak, aby zlepšila dostupnosť dokumentov širokej verejnosti prostredníctvom používateľsky prívetivého a intuitívneho rozhrania, ktoré umožňuje flexibilné prezeranie dokumentov a zjednodušuje ich prehľadávanie.

POROVNANIE S TECHNOLÓGIOU OCR

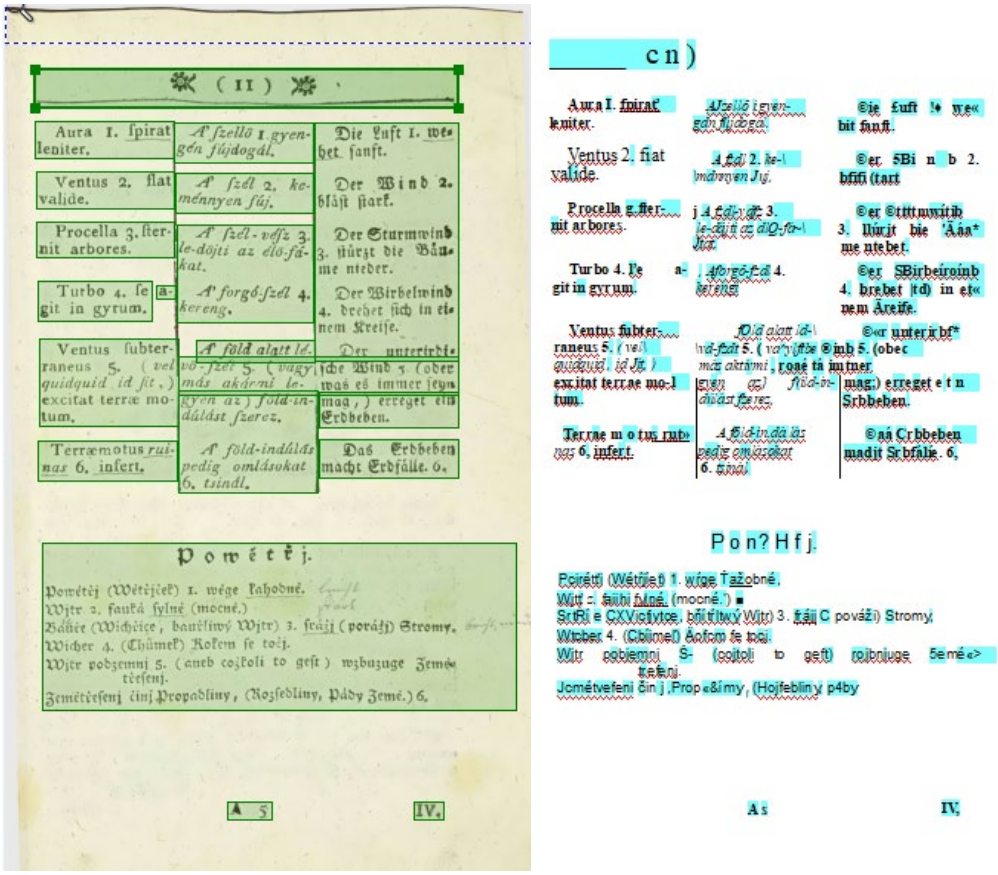
Jeden z riešiteľov projektu Skriptor dosiahol uspokojivé výsledky s nástrojom na optické rozpoznávanie textu *ABBYY FineReader* pri prepise strojopisných dokumentov. Prepísaný text vykazoval veľmi nízku chybovosť a bolo možné s ním prakticky okamžite pracovať. Najväčším benefitom bol čas ušetrený na procese segmentácie dokumentu a manuálnu prípravu vzorky *Ground Truth* (Smida 2023). Vzhľadom na to, že sme pracovali s tlačenými dokumentmi, rozhodli sme sa – v širšom kontexte všetkých skúmaných zbierok v projekte – porovnať kvalitu prepisu dokumentu v Transkribuse so softvérom *ABBYY FineReader*.

Kvalita segmentácie *Orbis Pictus* bola prijateľná a štruktúra textu bola vo väčšine prípadov zachovaná. Automatický prepis bol uspokojivý len vo fonte antikva a kurzíva, fraktúra a švabach vykazovali veľkú chybovosť, čo potvrdilo, že technológia OCR je určená skôr na prácu s jednoduchými fontami a dokumentmi prevažne z 20. storočia a novšími. Ani v jednom fonte softvér nedokázal spracovať prepis špeciálnych znakov.

Kvalita segmentácie a prepisu *Adparatu* bola vzhľadom na použité fonty oveľa uspokojivejšia. OCR si lepšie poradilo s kapitálkami, na ktorých prepis nebol model 11 dostatočne vytrénovaný. Aj v tomto prípade mal softvér problém s prepisom ligatúr.

ZÁVER

Digitálne humanitné vedy zahŕňajú kľúčové poznatky z jazykov a literatúry, histórie, hudby, médií a komunikácie, informatiky a informačných štúdií a spájajú tieto odlišné disciplíny do nových rámcov. V priebehu predchádzajúcich desaťročí sa na budovaní digitálnych humanitných vied,



Obr. 10 Ukážka prepisu Orbis Pictus pomocou OCR. Zdroj: ABBY FineReader

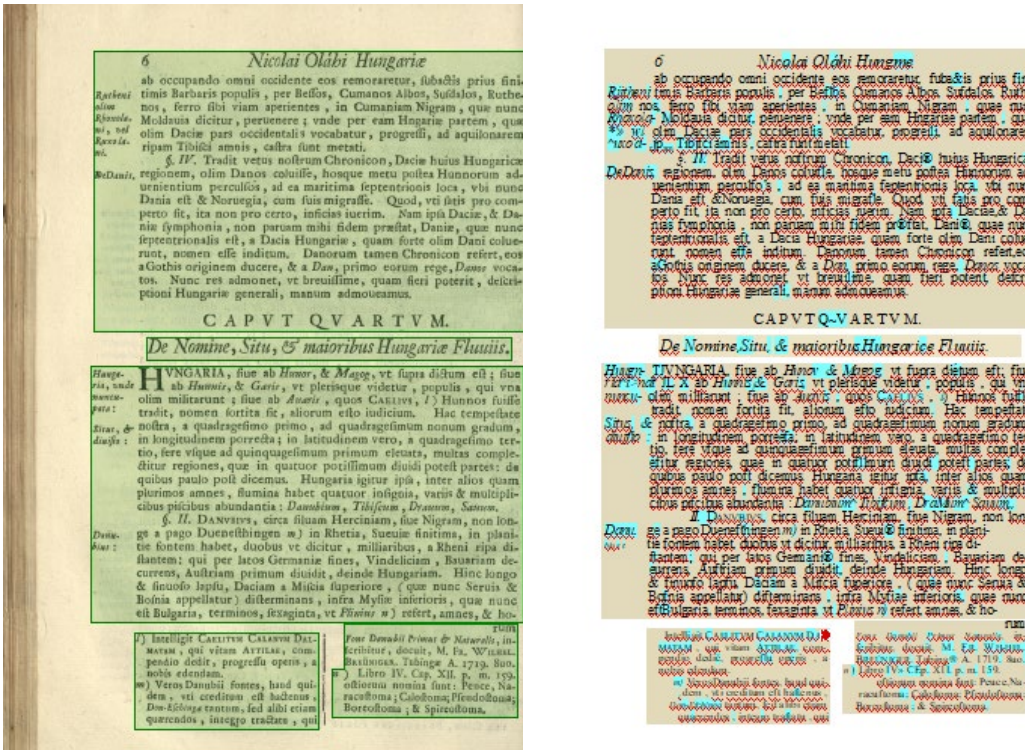
ako ich poznáme dnes, podieľala celá škála jednotlivcov, projektov a organizácií, ktorí digitálne technológie vnímali ako rozšírenie nástrojov tradičnej vedy a možnosť vytvárať a sprístupňovať archívy a databázy širšiemu okruhu používateľov (Burdick et al. 2012).

Projekt SKRIPTOR umožnil riešiteľom kompetentne participovať na výskumných aktivitách a svojimi výsledkami sa podieľať na udržateľnosti výsledkov výskumu európskeho projektu READ. Výsledky trojročnej práce s platformou založenou na technológii HTR+ môžu byť pozitívnou motiváciou pre širšiu odbornú komunitu, aby využila tento nástroj a všetky benefity, ktoré v oblasti uchovávania, využívania, sprístupňovania a prezentácie písomného kultúrneho dedičstva ponúka. Automatická transkripcia so slušnou presnosťou uľahčuje editovanie textu a jeho úpravy a transkribovaný originálny text (napríklad v latinčine, maďarčine, nemčine a i.) je možné ďalej automatizovane, pomocou kvalitných prekladačov, preložiť do iného jazyka. Tým sa dosť podstatne mení charakter práce archivárov a historikov. Na tomto mieste však treba zdôrazniť, že Transkribus nemá zámer ani ambíciu nahrádzať odbornú a vedeckú erudíciu historikov a archivárov. Automatická transkripcia je len jeden z krokov vedeckej práce, po ktorom prirodzene nasleduje historický výskum textu a kontextu transkribovaných dokumentov, editácia textov získaných transkripciou, identifikácia entít, kľúčových slov (dátumov, mien osôb, geografických názvov, korporácií a pod.) a napokon ich sprístupnenie na využitie v širšom meradle. Ťažkú orientáciu v rukopisných zbierkach a tlačiach s rozsahom stoviek strán či prácne vyhľadávanie v dobových, historických archívnych pomôckach, v ktorých sa skrýva množstvo dodnes neodhalených informácií, možno za relatívne krátky čas nahradiť rýchlym a dôsledným prehľadávaním textu pomocou klávesových skratiek.

V rámci vzdelávacích aktivít pripravili riešitelia rozsiahlu a podrobnú príručku na prácu s verziou *Transkribus Expert Client* a verziou *Transkribus Lite*. Organizovali a v súvislosti s udržateľnosťou projektu budú aj v nasledujúcich mesiacoch organizovať praktické školenia na prácu s platformou Transkribus.

POĎAKOVANIE

Poďakovanie autoriek patrí vedúcemu Odboru knižnično-informačných služieb CVTI Mgr. Tomášovi Fialovi a zamestnancom Digitalizačného pracoviska CVTI Ing. Branislavovi Jandorfovi a p. Martinovi Krajčírovi za zdigitalizovanie



Obr. 11 Ukážka prepisu Adparatu pomocou OCR. Zdroj: ABBY FineReader

vyše sedemstostranového diela Mateja Bela *Adparatvs ad Historiam Hvngariae, sive collectio miscella, Monumentorum ineditorum partim, partim editorum, sed fugientium. Conquisiut, in Decades partitus est, & Praefationibus, atque Notis illustrauit* (1735) a takmer dvestostranovej učebnice Jána Amosa Komenského *Orbis Pictus* (1798), a za precíznu a citlivú prácu s oboma historickými dokumentmi.

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

POUŽITÉ ZDROJE

Baselines models, ([2023]). Online. In: Transkribus: Help Center. Dostupné na: https://help.transkribus.org/baselines-models. [dátum citovania 2024-04-10].

BASUMALLICK, Chiradeep, (2023). *What Is OCR (Optical Character Recognition)? Meaning, Working, and Software*. Online. In: Spiceworks. 2006 – 2024. Dostupné na: https://www.spiceworks.com/tech/tech-general/articles/optical-character-recognition-ocr. [dátum citovania 2024-04-10].

BEL, Matthias, (1735). *Adparatvs ad Hvngariae, sive collectio miscella, Monumentorum ineditorum partim, partim editorum, sed fugientium. Conquisiut, in Decades partitus est, & Praefationibus, atque Notis illustrauit, Matthias Bel.* Online. Posonii: Typis Joannis Paulii Weber. Dostupné na: https://repo.umb.sk/handle/123456789/493. [dátum citovania 2024-04-25].

BERRY, David M., (2022). *What are the digital humanities?* Online. In: The British Academy. Dostupné na: https://www.thebritishacademy.ac.uk/blog/what-are-digital-humanities. [dátum citovania 2024-04-10]

BURDICK, Anne; DRUCKER, Johanna; LUNENFELD, Peter; PRESNER, Todd a SCHNAPP, Jeffrey, (2012). *Digital Humanities*. Online. MIT Press. ISBN 978-0-262-01847-0. Dostupné na: https://mitpress.mit.edu/9780262528863/digital_humanities. [dátum citovania 2024-04-10].

COMENII, Ioann Amos, (1798). *Joann. Amos Comenii Orbis pictus, in hungaricum, germanicum et slavicum translatus et hic ibive emendatus*. Online. Posonii: Sumtibus & Typis Simonis Petri Weber. Dostupné na: https://repo.umb.sk/handle/123456789/492. [dátum citovania 2024-04-25].

Data preparation, ([2023]). Online. In: Transkribus: Help Center. Dostupné na: https://help.transkribus.org/data-preparation. [dátum citovania 2024-04-25].

DIETRICH, Felix, (2023). *OCR vs. HTR or "What is AI, actually?"*. Online. In: READ-COOP SCE. Dostupné na: https://readcoop.eu/insights/ocr-vs-htr. [dátum citovania 2024-04-10].

GOLD, Matthew K. a KLEIN, Lauren F., (2016). Digital Humanities: The Expanded Field. Online. In: GOLD, Matthew K. a KLEIN, Lauren F. (ed.). *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press. ISBN 978-1-4529-5149-2. Dostupné na: https://dhdebates.gc.cuny.edu/read/untitled/section/4960b2c5-e2da-4a8b-8e07-ca1a791b963c#toc. [dátum citovania 2024-04-10].

Horizon Impact Award 2020: 10 finalists short-listed, (2020). Online. In: DIRECTORATE-GENERAL FOR RESEARCH AND INNOVATION. Research and innovation. Dostupné na: https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/horizon-impact-award-2020-10-finalists-short-listed-2020-09-15_en. [dátum citovania 2024-04-10].

Character error rate and learning curve, ([2023]). Online. In: Transkribus: Help Center. Dostupné na: https://help.transkribus.org/character-error-rate-and-learning-curve. [dátum citovania 2024-04-25].

KATUŠČÁK, Dušan, (2020). Digital Humanities a automatická transkripcia rukopisných textov. Online. *ITlib: informačné technológie a knižnice*. č. 1, s. 6 – 16. Dostupné na: https://itlib.cvtisr.sk/wp-content/uploads/docs//2020/1_2020/3.pdf. [dátum citovania 2024-04-10].

KATUŠČÁK, Dušan a NAGY, Imrich (ed.), (2023). *Automatická transkripcia historických dokumentov – metodická príručka na prácu s platformou Transkribus*. Online. Banská Bystrica: Belianum, Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici. ISBN 978-80-557-2070-8. Dostupné na: https://doi.org/10.24040/2023.9788055720708. [dátum citovania 2024-04-10].

KEIJSER, L. a MOSS, A. D. E., (2019). *Transkribus Transcription Convention Project De ijsberg zichtbaar*. Online. In: NATIONAAL ARCHIEF. Kennisnetwerk Informatie en Archief. Dostupné na: https://kia.pleio.nl/attachment/entity/387ab05a-d087-42f3-8b07-cbc56380dead. [dátum citovania 2024-04-25].

KETTUNEN, Kimmo; KESKUSTALO, Heikki; KUMPULAINEN, Sanna; RÄÄKKÖNEN, Tuula a RAUTIAINEN, Juha, (2023). Optical character recognition quality affects subjective user perception of historical newspaper clippings. Online. *Journal of Documentation*. vol. 79, no. 7, s. 137 – 156. ISSN 0022-0418. Dostupné na: https://doi.org/https://doi.org/10.1108/JD-01-2023-0002. [dátum citovania 2024-04-10].

MISTRÍK, Jozef, (1993). Prepis z iného jazyka. In: *Encyklopédia jazykovedy*. Bratislava: Obzor, 1993, s. 343. ISBN 80-215-0250-9.

New Document Editor, (2024). Online. In: READ-COOP SCE. Dostupné na: https://readcoop.eu/new-document-editor. [dátum citovania 2024-04-10].

NIŽNÍKOVÁ, Lucia a MIKUŠKOVÁ, Michaela, (2022). Modely automatickej transkripcie štvorjazyčného diela J. A. Komenského Orbis Pictus (1798). Online. In: KATUŠČÁK, Dušan a NAGY, Imrich (ed.). *Automatická transkripcia slovacikálnych historických dokumentov*. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2022, s. 168 – 192. ISBN 978-80-557-2020-3. Dostupné na: https://doi.org/10.24040/2022.9788055720203. [dátum citovania 2024-04-10].

Recognition and Enrichment of Archival Documents, (2015). Online. In: CORDIS: EU research results. Dostupné na: https://doi.org/10.3030/674943 [dátum citovania 2024-04-10].

SMIDA, Matej, (2023). Možnosti automatickej transkripcie v platforme Transkribus na príklade správ o vybavovaní sťažností občanov v období komunistickej diktatúry. Online. *Acta Historica Neosoliensia*. roč. 26, č. 1, s. 125 – 148. ISSN 2453-7845. Dostupné na: https://doi.org/10.24040/ahn.2023.26.01.125-148. [dátum citovania 2024-04-25].

Transkribus Sites – publishing documents made easy, (2024). Online. In: READ-COOP SCE. Dostupné na: https://readcoop.eu/transkribus-sites-publishing-documents-made-easy. [dátum citovania 2024-04-10].

Using Transkribus for OCR with printed books, (2023). Online. In: READ-COOP SCE. Dostupné na: https://readcoop.eu/using-transkribus-for-ocr-with-printed-books/. [dátum citovania 2024-04-10].

■ **Článok bol recenzovaný.**