AUTOMATICKÁ TRANSKRIPCIA HISTORICKÝCH DOKUMENTOV

metodická príručka na prácu s platformou Transkribus

Dušan Katuščák – Imrich Nagy
(eds.)

2023

Automatická transkripcia historických dokumentov

metodická príručka na prácu s platformou Transkribus

Mária Bôbová, Dušan Katuščák, Alica Kurhajcová, Patrik Kunec, Pavol Maliniak,
Michaela Mikušková, Imrich Nagy, Lucia Nižníková, Oto Tomeček

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

Editori: Dušan Katuščák, Imrich Nagy

Autori: © Mária Bôbová, Dušan Katuščák, Alica Kurhajcová, Pavol Maliniak, Michaela Mikušková, Imrich Nagy, Lucia Nižníková, Patrik Kunec, Oto Tomeček

Jazyková korektúra: Lucia Nižníková

Grafická úprava: Miroslav Chladný

Verzia Transkribus Expert Client 1.26.0

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

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

DOI: https://doi.org/10.24040/2023.9788055720708

Zdroj fotografie na obálke: https://readcoop.eu/national-archives-finland-takes-first-steps-towards-handwritten-text-recognition/

ISBN 978-80-557-2070-8

Slovo na úvod

Technologický pokrok vo využívaní nástrojov strojového učenia (*Machine Learning*) a umelej inteligencie AI (*Artificial Intelligence*) sa postupne stáva súčasťou našej každodennosti a vedome či nevedome sme s ním konfrontovaní aj pri rôznych špecifických odborných činnostiach, pri ktorých bolo nahradenie vedomostí a zručností človeka strojom donedávna nepredstaviteľné. Neznamená to však, že by sa odbornosť človeka stávala zbytočnou. Sú to práve invenčnosť, zručnosť a um človeka, ktoré využili a adaptovali existujúce i novo sa vyvíjajúce technológie na zvládanie presne definovateľných, rutinných a opakujúcich sa algoritmov. Takým procesom je aj zostavovanie textov na ľubovoľné témy prostredníctvom chatovacích robotov, ktoré je v súčasnosti až emblematickým symbolom pokroku vo využiteľnosti AI.

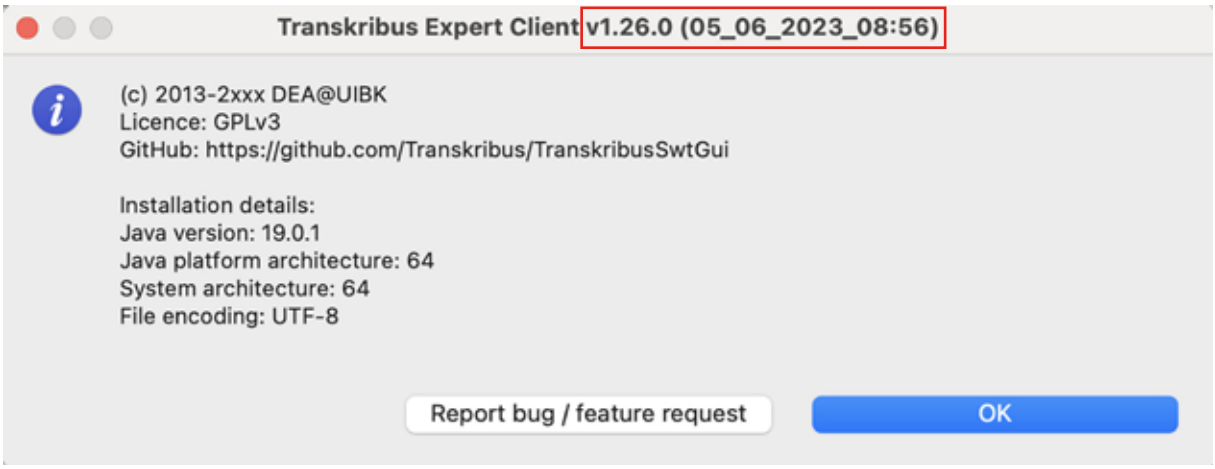
Tak trochu v tieni týchto populárnych nástrojov zostávajú dlhodobo vyvíjané a v praxi overované nástroje AI schopné vskutku mimoriadnym spôsobom zmeniť a v podstate nanovo zadefinovať vysoko odborné činnosti v jednotlivých profesiách. Pozoruhodným príkladom toho je aj platforma Transkribus vyvinutá v multilaterálnej spolupráci významných európskych vedeckých inštitúcií v rámci projektov *tranScriptorium* (2013 – 2015) a *READ* (2016 – 2019) financovaných z programov EÚ. Lídrom tejto spolupráce je Univerzita v Innsbrucku a vedúcou postavou Dr. Günter Mühlberger, ktorí výsledky predchádzajúcich projektov pretavili do veľmi dynamicky sa vyvíjajúceho a účinného nástroja na automatickú transkripciu dokumentov v rukopisnej aj tlačenej podobe ľubovoľnej geografickej, historickej či jazykovej proveniencie. Vďaka tomu je v súčasnosti Transkribus verejne dostupný komerčný produkt, ktorý prostredníctvom združenia READ-COOP European Cooperative Society ponúka všetkým individuálnym a inštitucionálnym záujemcom riešenie pre vskutku efektívnu digitalizáciu historických dokumentov s plnotextovými digitálnymi výstupmi v najrozličnejších formátoch podľa požiadavky zadávateľa. Pre pamäťové inštitúcie a ich používateľov z radov laickej i odbornej verejnosti je to doslova revolučná zmena, ktorá zásadným spôsobom do budúcnosti zmení ich prácu a má potenciál priniesť mimoriadne výsledky v poznaní a sprístupňovaní našej histórie a nehmotného kultúrneho dedičstva.

Univerzita Mateja Bela v Banskej Bystrici sa v spolupráci so Štátnou vedeckou knižnicou v Banskej Bystrici v rámci projektu *SKRIPTOR – Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov podporeného Agentúrou pre podporu vedy a výskumu (APVV-19-0456)* podujala aplikovať nástroj Transkribus na slovacikálne historické rukopisné a tlačené dokumenty a overiť jeho využiteľnosť v podmienkach slovenských pamäťových inštitúcií (archívov). Na Slovensku ide o ojedinelý pilotný projekt, ktorý môže otvoriť cestu k žiaducemu nasadeniu moderných technológií pri digitalizačných projektoch našich pamäťových inštitúcií v záujme širokého sprístupnenia informácií z digitalizovaných dokumentov a ich ďalšieho odborného využitia.

V rámci riešenia projektu sa už podarilo vytvoriť desiatku funkčných modelov na automatickú transkripciu slovacikálnych rukopisov zo 16. – 20. storočia a tiež historických tlačí. Významnou pridanou hodnotou je nadobudnutie know-how práce na platforme Transkribus. Jeho sprostredkovanie záujemcom a zástupcom pamäťových inštitúcií zo Slovenska formou workshopov považujeme za zmysluplné zavŕšenie našej práce. Za týmto účelom sme zostavili metodickú príručku, ktorá v postupných krokoch predstavuje jednotlivé fázy digitalizácie a automatickej transkripcie dokumentu na platforme Transkribus.

Na tomto mieste je dôležité upozorniť, že platforma Transkribus sa stále vyvíja. Na stránkach https://readcoop.eu/transkribus/ sa nachádzajú manuály a videá na prácu s Transkribom. Niektoré inštrukcie a názorné ukážky, ktoré boli aktuálne v minulosti a podľa ktorých boli inštrukcie

pripravené, už neodrážajú vlastnosti a funkcie nových verzií. Metodickú príručku sme pripravili podľa poslednej verzie Transkribus Expert Client 1.26.0 z 5. júna 2023. S touto verziou by mal pracovať aj užívateľ tejto príručky, ktorú budeme v budúcnosti podľa potreby aktualizovať.



Obrázok 1 Transkribus Expert Client verzia 1.26.0

Veríme, že využitie možností AI vo forme práce s Transkribom prinesie novú dynamiku do digitalizácie, uchovávania a sprostredkovania nehmotného kultúrneho dedičstva na Slovensku.

Dušan Katuščák – Imrich Nagy

1 Registrácia a účet na platforme Transkribus

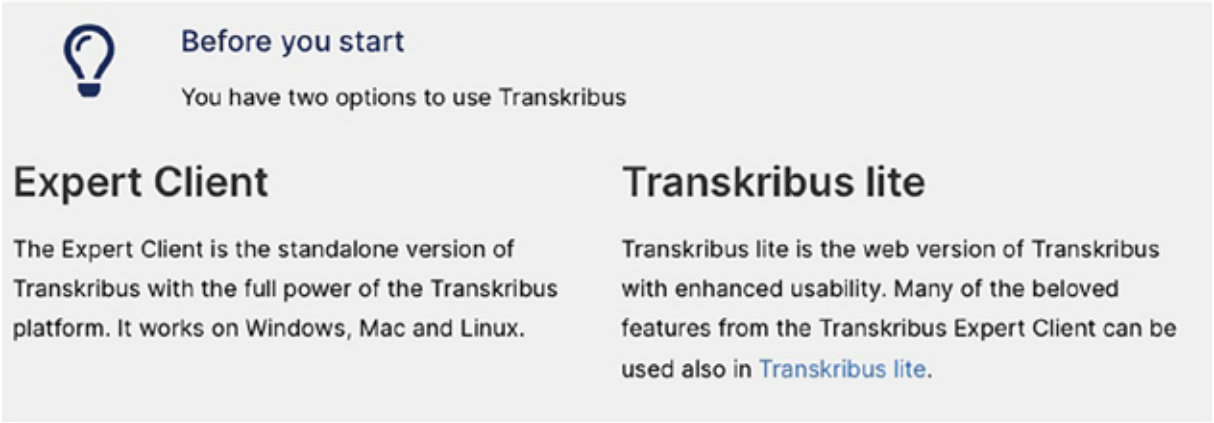
1.1 Pripojenie na internet

Na prácu na platforme Transkribus je potrebné mať k dispozícii nepretržité vysokorýchlostné pripojenie na internet. Všetky vaše úkony sa budú robiť v režime vzdialeného prístupu na serveroch platformy Transkribus. Všetky súbory a verzie strán, s ktorými budete pracovať, sa budú ukladať na serveroch platformy. Výhodou je, že k nim budete mať prístup z ktoréhokoľvek miesta a ktoréhokoľvek počítača (cez webový prehliadač alebo prostredníctvom nainštalovanej aplikácie).

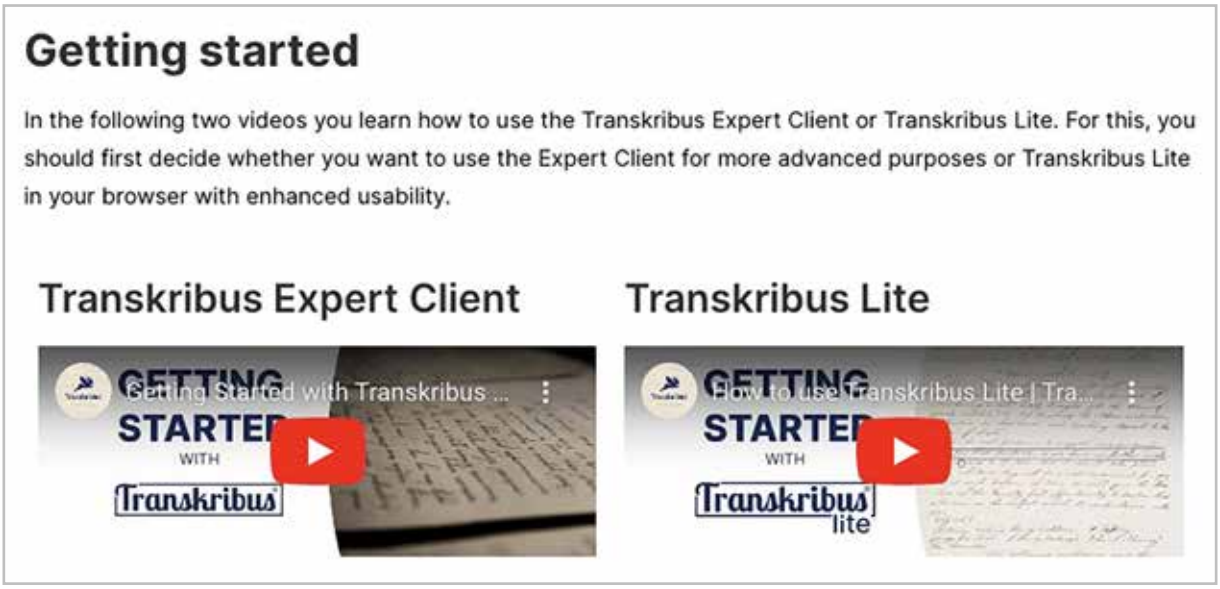
1.2 Voľba rozhrania pre prácu na platforme Transkribus

Používateľ platformy Transkribus má dve možnosti:

1. používať odborného (expertného) klienta *Transkribus Expert Client* (ďalej aj Transkribus expert klient alebo expert klient),
2. používať webového klienta cez internetový prehliadač *Transkribus Lite*.



Obrázok 2 Úvodné informácie o dvoch verziách platformy Transkribus



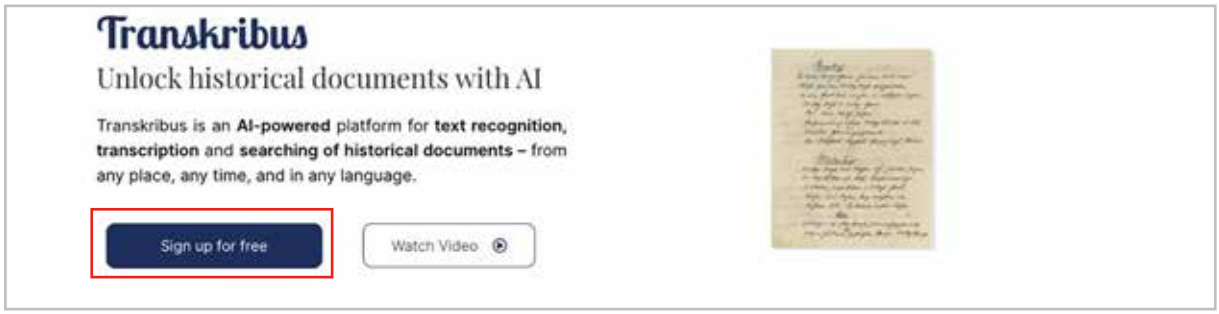
Obrázok 3 Videonávody pre dve možnosti použitia platformy Transkribus

1.3 Registrácia

Na prácu s platformou Transkribus sa musíte zaregistrovať. Ak už máte svoj Transkribus účet, prejdite na ďalší krok.

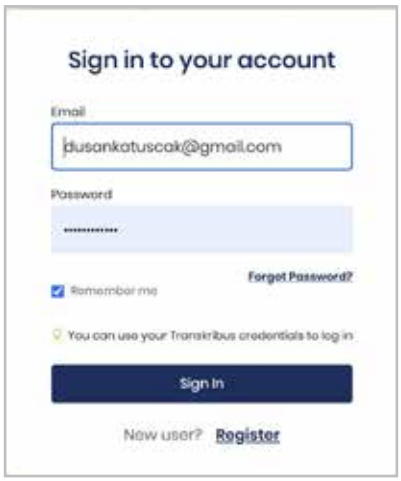
Postup registrácie na platforme Transkribus:

1. Otvorte webstránku https://readcoop.eu/transkribus/
2. Kliknite na voľbu **Prihlásiť sa zdarma** (*Sign up for free*).
3. Vyplňte registračné údaje. Platia pre všetky prístupy na platformu Transkribus, teda do Transkribus expert klienta, Transkribus Lite a i.
4. Kliknite na tlačidlo **Registrovať sa** (*Register*).
5. Registráciu overte kliknutím na odkaz v e-maili, ktorý dostanete do e-mailovej schránky, s ktorou ste sa registrovali.



Obrázok 4 Vytvorenie účtu cez voľbu Prihlásiť sa zdarma

Obrázok 5 Registrácia. Okno na zápis registračných údajov používateľa na platformu Transkribus



Obrázok 6 Prihlásenie do účtu Transkribus

Prihláste sa do Transkribu pomocou e-mailovej adresy a hesla zvoleného počas procesu registrácie. Vaše údaje platia na používanie platformy stále, preto ich nemeňte.

1.4 Osobný účet

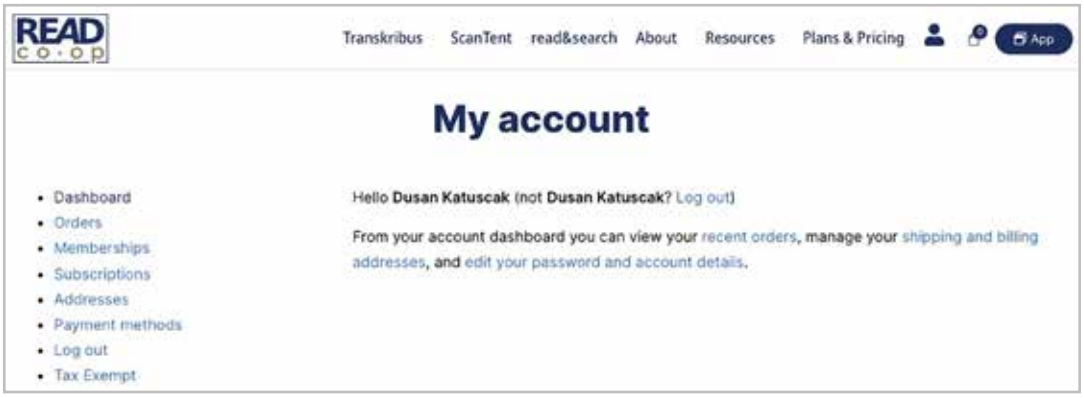
Registrácia účtu Transkribus je jednoduchá, stačí sa bezplatne zaregistrovať. Ak ste sa už registrovali, máte vytvorený vlastný účet a s ním máte prístup:

1. na platformu Transkribus,
2. do odborného (expert) klienta a Transkribus Lite vo webovom prehliadači (napr. Google Chrome, Safari, Edge a i.).

Transkribus Lite nie je potrebné inštalovať. Je dostupný cez internetový prehliadač (napr. Chrome, Edge, Safari; odporúčame Chrome).

Transkribus expert klient sa musí nainštalovať. Pre odborníkov obsahuje viac funkcií. Inštrukcie k inštalácii sa nachádzajú v ďalšej časti tejto metodickej príručky.

1.4.1 Prístup do účtu



Obrázok 7 Prístup do účtu platformy Transkribus

Kliknite na ikonku postavy vpravo hore. Zobrazí sa váš osobný účet. V účte je v položke **Objednávky** (*Orders*) história platieb, napr. platba za účasť na výročnej konferencii Transkribus a pod.

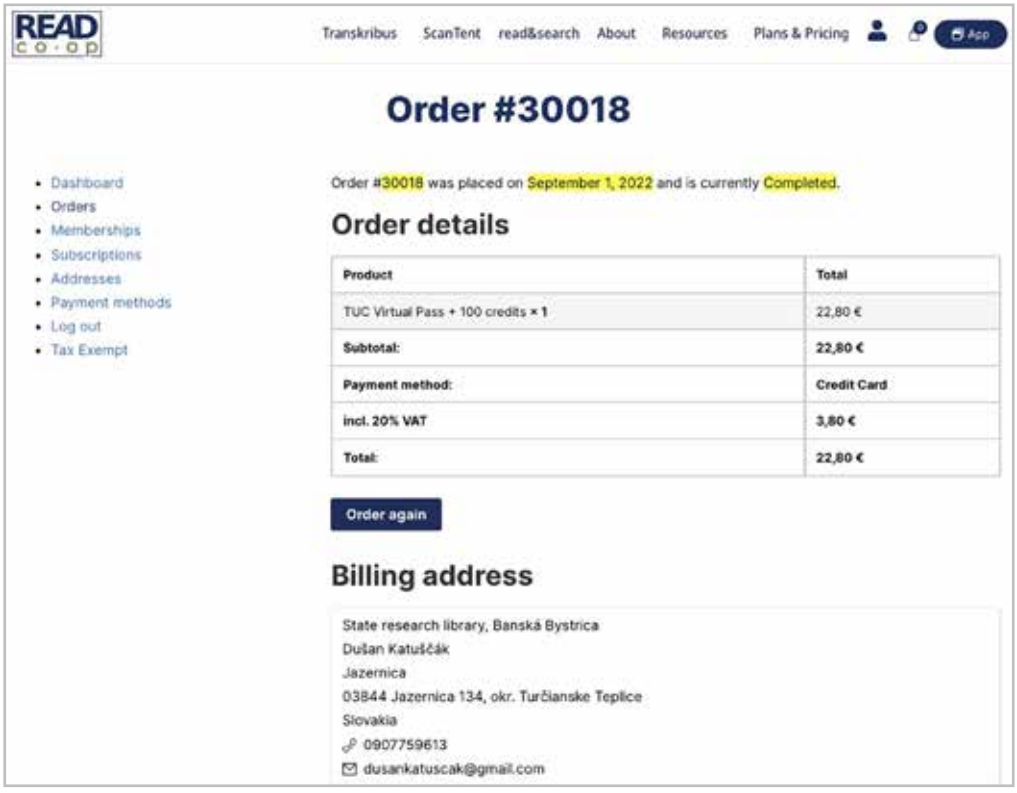
1.4.2 Správa účtu

Po registrácii môžete svoj účet spravovať zo stránky účtu.

Všetky dokumenty nahraté na platformu Transkribus sú súkromné, čo znamená, že nikto okrem vás k nim nemá prístup. Ak pracujete na projekte digitalizácie a transkripcie historických dokumentov s viacerými odborníkmi, môžete ich pridať k svojej zbierke a po dohode s nimi im prideliť oprávnenia.

Poznámka: Na platforme sú plne podporované všetky smernice EÚ o ochrane osobných údajov a súkromia.

Obrázok 8 Účet s položkou Objednávky



Obrázok 9 Účet s rozpisom položky Objednávky

V účte si môžete prezerať nedávne objednávky, vidieť svoje členstvo v združení READ-COOP a aktuálne odbery, spravovať adresy alebo upravovať podrobnosti o svojom účte.

Po registrácii má každý používateľ bezplatne k dispozícii 500 kreditov, ďalšie kredity si môže dokupovať podľa potreby.

2 Transkribus expert klient

Transkribus expert klient je softvér, ktorý sa inštaluje na osobný počítač. Samotná výpočtová platforma Transkribus je nainštalovaná na vzdialenom serveri. Informačná architektúra platformy v tomto prípade je klient – server.

So zbierkami a dokumentmi pracujete v expert klientovi. Všetky zbierky a dokumenty a ich verzie sú umiestnené a dostupné cez váš účet na serveri.

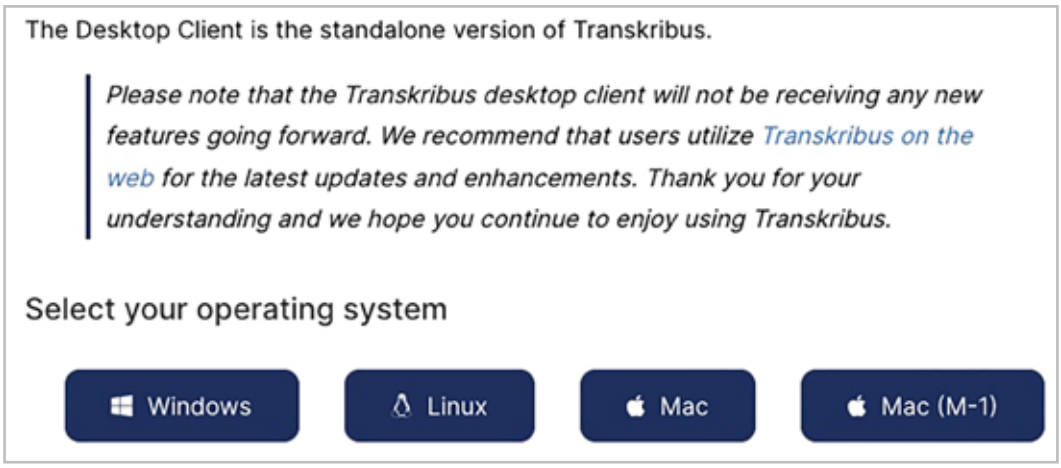
Expert klient je desktopový klient a samostatná verzia Transkribu. Desktopový klient Transkribus nebude podľa vyjadrení vývojového tímu v júli 2023 v budúcnosti rozširovaný o žiadne nové funkcie. Kým bude verzia Transkribus Lite plne funkčná, zoznamujeme záujemcov o prácu s platformou Transkribus s verziou expert klient, v ktorej sú dostupné mnohé funkcie, ktoré ešte nie sú v Transkribus Lite. Predpokladáme, že Transkribus Lite s komplexnejšími funkciami bude k dispozícii v roku 2024.

Odporúčame, aby používatelia využívali Transkribus Lite https://lite.transkribus.eu/ s najnovšími aktualizáciami a vylepšeniami. V júli 2023 bol Transkribus Lite limitovaný veľkosťou nahrávaného súboru PNG a JPG do 10 MB. Funkcie z expert klienta sa v súčasnosti presúvajú do verzie Transkribus Lite.

Obrázok 10 Limit veľkosti súboru v Transkribus Lite

2.1 Inštalácia Transkribus expert klienta

Inštaláciu zvoľte podľa toho, aký operačný systém máte vo svojom osobnom počítači. Expert klienta môžete inštalovať na počítač s operačnými systémami Windows, Linux, Mac, Mac M-1.



Obrázok 11 Inštalácia podľa operačného systému

Poznámka: Aby ste mohli pracovať s Transkribom, musíte mať nainštalovaný programovací jazyk Java. Najnovšiu verziu si môžete stiahnuť z oficiálnej stránky Oracle tu https://www.oracle.com/java/technologies/downloads/

Po registrácii účtu si môžete bezplatne stiahnuť Transkribus z domovskej stránky. Zatiaľ čo všetky ostatné funkcie Transkribu je možné využívať bezplatne, na automatické prepisy sú potrebné kredity. Všetky informácie o kreditnom systéme nájdete v kapitole *6 Priebeh automatickej transkripcie v platforme Transkribus*.

Pripomíname, že každý nový účet Transkribus dostane na testovanie 500 kreditov zdarma. Zámerom je takýmto spôsobom podporiť tých, ktorí majú malé transkripčné projekty, ktorým stačí rozpoznať len zopár strán.

Kliknite na tlačidlo operačného systému, ktorý máte v počítači. Začne sa sťahovanie.

Po kliknutí na tlačidlo sťahovania na domovskej stránke READ-COOP získate súbor ZIP, ktorý je potrebné rozbaliť (kliknite pravým tlačidlom myši na priečinok a vyberte možnosť *Rozbaliť všetko*).

Otvorte na svojom počítači adresár Transkribus. Tu nájdete spustiteľné súbory pre váš operačný systém.

Ak máte operačný systém Windows, dvakrát kliknite na súbor **.exe**. Ak máte operačný systém Mac kliknite na **.command**. Ak máte Linux, kliknite na **.sh**.

Technické pokyny:

Ak je operačný systém založený na Ubuntu 17.04, je potrebná inštalácia libwebkit: sudo apt install libwebkitgtk-1.0-0

Ak nemáte práva správcu, systém Windows zobrazí varovné hlásenie, napríklad *Váš počítač je chránený systémom Windows*, atď.

Nepotvrdzujte, namiesto toho zvoľte *Viac informácií* a potvrďte, že chcete Transkribus aj tak spustiť.

Ak program spúšťaťe prvýkrát, nemusí sa spustiť, pretože ide o nepodpísanú aplikáciu (správa ... *nie je možné otvoriť, pretože je od neidentifikovaného vývojára*). V takom prípade kliknite pravým tlačidlom myši (alebo kliknite so stlačeným klávesom *Control*) na aplikáciu a vyberte možnosť *Otvoriť*. V zobrazenom dialógovom okne znova kliknite na *Otvoriť*. Prípadne kliknite pravým tlačidlom myši na *Track Pad*, aby ste otvorili kontextové menu a pridali bezpečnostnú výnimku pre Transkribus.

Ďalšia možnosť: kliknite pravým tlačidlom myši na ikonu programu -> Otvoriť (v kontextovom menu) -> cez Terminál. Ak sa aplikácia vôbec nespustí, môžete sa pokúsiť presunúť aplikáciu z priečinka Download, tj skopírovať alebo presunúť do iného cieľového priečinka, ako je Pracovná plocha.

Chybové hlásenie pri pokuse o spustenie aplikácie z terminálu pomocou open -a Transkribus.app je *LSOpenURLsWithRole() zlyhalo pre aplikáciu /Users/xxx/Desktop/Transkribus.app s chybou -10810.*

Alternatívnym riešením na spustenie programu je spustenie nového terminálu (hľadajte t*erminál* po stlačení cmd + medzera), potom cd do adresára, kde bol rozbalený Transkribus, napr. adresár *Downloads*: Potom spustite program priamo zo štartovacieho skriptu, ktorý je súčasťou balíka Transkribus.app: /Transkribus.app/Contents/MacOS/Transkribus

Transkribus je obsiahnutý v hlavnom jar súbore Transkribus-<verzia>.jar

Ak chcete spustiť program z príkazového riadka, zadajte java -jar Transkribus-<verzia>.jar

Niektoré problémy sa môžu vyskytnúť, ak je 32-bitová verzia Java nainštalovaná na 64-bitovom operačnom systéme.

Poznámka: Ak chcete spustiť skripty v systéme Mac (alebo Linux), možno bude potrebné, aby boli spustiteľné z príkazového riadka (akákoľvek verzia pred 0.6.8).

Základy konzoly Mac zmeňte do priečinka programu pomocou príkazov „cd“; chmod +x Transkribus.command (alebo chmod +x Transkribus.sh pre Linux).

Okrem toho v balíku Transkribus nájdete niekoľko súborov skopírovaných do vášho počítača. *config.properties* je možné použiť na úpravu jednoduchých vlastností vzhľadu; súbor *virtualKeyboards.xml* je možné pou-

žiť na určenie sady virtuálnych klávesníc; *logback.xml* je možné použiť na úpravu vlastností protokolovania (len pre skúsených používateľov). Podpriečinok *libs* obsahuje potrebné knižnice pre všetky platformy.

V súčasnosti sú podporované Windows 32/64 bit; Linux 32/64 bit; OSX 64 bit. Ak sa stále zobrazuje chybové hlásenie *Prihlásenie zlyhalo: už pripojené*, problém môže byť v proxy serveri. Po spustení programu kliknite na tlačidlo domovskej ponuky vľavo hore a vyberte *Nastavenia servera proxy.* V nasledujúcom dialógovom okne môžete nastaviť hostiteľa proxy, port, meno používateľa (voliteľné) a heslo (voliteľné). Toto je odporúčaný spôsob používania proxy servera. Prípadne môžete upraviť spúšťací skript (napr. Transkribus.bat na Windows, Transkribus.sh na Linuxe), aby obsahoval premenné prostredia pre proxy server:

Prihlásenie na server nie je možné cez Transkribus, ale na stránke to funguje.

Ďalší možný dôvod chybového hlásenia *Už pripojené*: znamená, že vaša inštalácia Java môže byť zastaraná a nemôže vytvoriť zabezpečené pripojenie k serveru. Svoju nainštalovanú verziu môžete skontrolovať otvorením terminálu/príkazového riadku a zadaním *java-version*. Ak narazíte na tento problém, skúste aktualizovať Javu na počítači. Odporúčame aktuálnu verziu nie staršiu ako Java 11 (Oracle alebo OpenJdk). Verzia Transkribus pre Mac obsahuje runtime Java. Ak narazíte na tento problém na Macu, stiahnite si nový balík z https://readcoop.eu/ a aktualizujte inštaláciu. Ak chyba pretrváva, kontaktujte info@readcoop.eu ideálne vrátane log súboru vašej inštalácie (z adresára Transkribus logs/TrpGui.log) a/alebo informácií o verzii Java a o operačnom systéme.

Poznámka: Verzia expertného klienta pre Mac sa dodáva s Javou dodávanou v rámci aplikácie. Ak je táto verzia Java zastaraná, môžete ju skúsiť odstrániť alebo nahradiť aktualizovanou verziou. Ak chcete nájsť súbory vo vyhľadávači Mac, kliknite pravým tlačidlom myši (alebo kliknite na cmd) na aplikáciu Transkribus *v zobrazení programov, v kontextovej ponuke kliknite na* Zobraziť obsah balíkov*, a potom prejdite do podpriečinku* Obsah/MacOS*. Tam podpriečinok* jre *obsahuje túto verziu Java. Ak odstránite tento priečinok, spúšťač aplikácie sa pokúsi nájsť Java vo vašom systéme.*

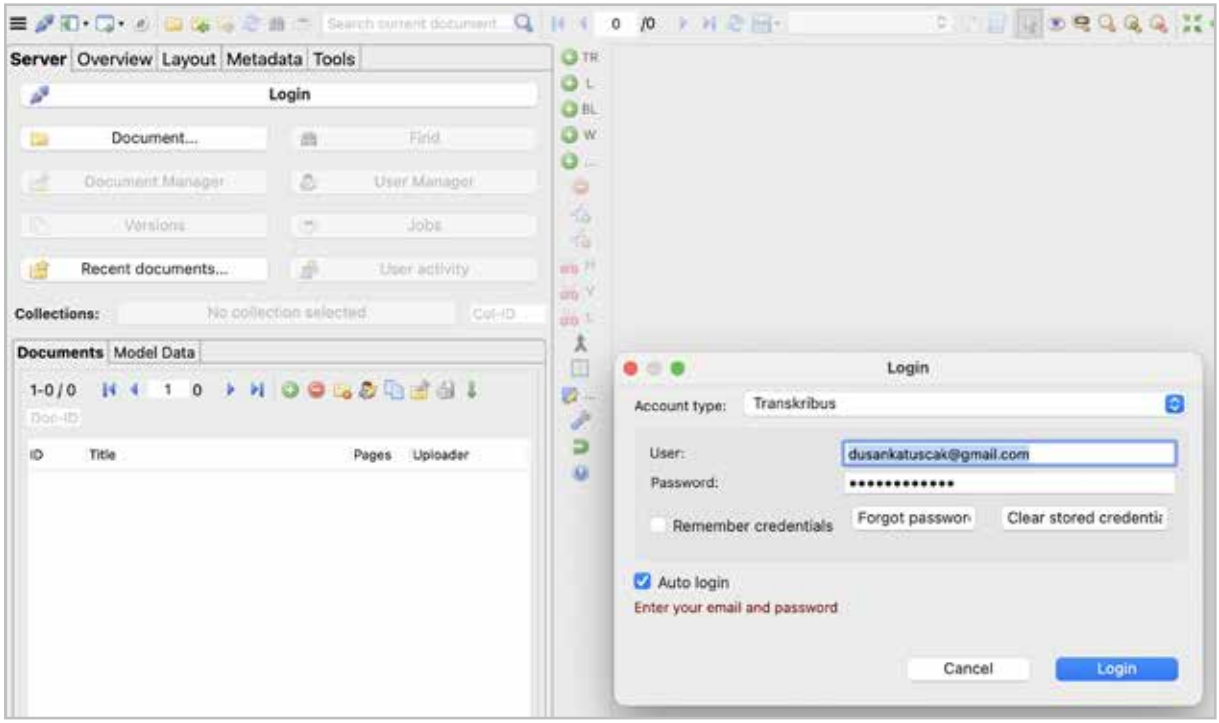
Možno bude potrebné nakonfigurovať proxy server cez hlavné menu *Nastavenia proxy*. Po otvorení príkazového súboru na Macu Transkribus oznámi, že je nainštalovaná nesprávna verzia Java. Je tu však nainštalovaná najnovšia verzia Java. Problém je v tom, že staršia verzia môže byť stále nainštalovaná a nastavená ako predvolená Java. Predvolenú verziu môžete skontrolovať otvorením terminálu a zadaním *java-version*. Ak chcete problém vyriešiť, môžete si stiahnuť najnovší jdk ako balík .tar.gz. Nastavte inštaláciu Java 11 (alebo jednu z nasledujúcich verzií) ako predvolenú v príkazovom riadku, napríklad podľa pokynov; alebo len skúste preinštalovať najnovšiu Java JDK z inštalátora na https://www.oracle.com/java/technologies/downloads/

Ak máte príliš málo RAM, skúste alokovať viac hlavnej pamäte otvorením Transkribus.bat a nastavte napr. java -Xmx2048m -jar Transkribus-1.14.0.jar. Spustite Transkribus s týmto .bat súborom.

Niektoré IT oddelenia blokujú port SSL 443 a/alebo neznáme aplikácie cez firewall. Ak je to tak, obráťte sa na svoje IT oddelenie.

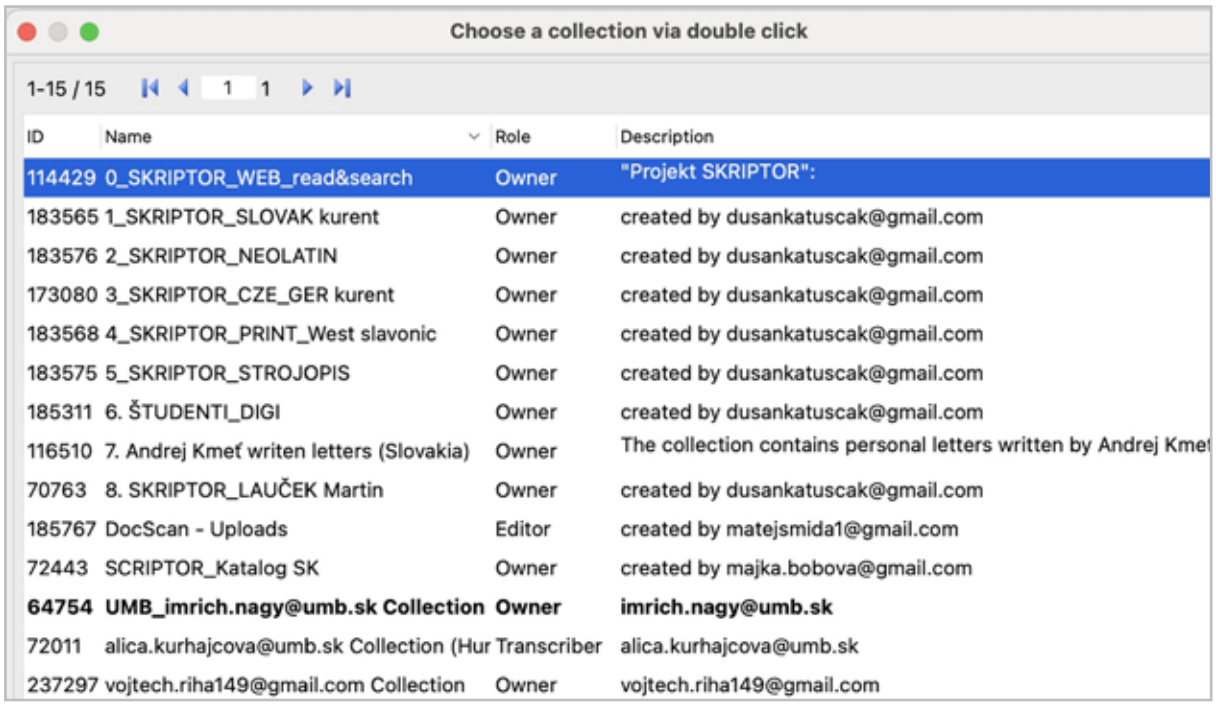
Tmavý režim na Macu môže niekedy spôsobiť problémy s Transkribom, takže ak Transkribus na vašom Macu nefunguje správne a máte zapnutý tmavý režim, skúste ho vypnúť. Po vypnutí tmavého režimu pravdepodobne budete musieť znova nainštalovať Transkribus, aby sa zmena prejavila.

2.2 Prihlásenie do Transkribus expert klienta



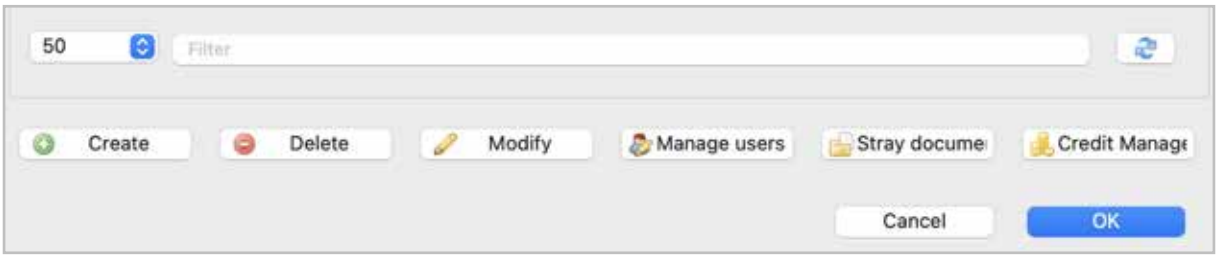
Obrázok 12 Hlavná pracovná plocha Transkribus expert klienta a prihlásenie

Po prihlásení do expert klienta máte prístup do svojho konta a k svojim súkromným zbierkam pomenovaným podľa vašej e-mailovej adresy. Na účely tejto metodiky a workshopu sme vytvorili pracovnú zbierku ID 190485.



Obrázok 13 Niektoré zbierky v expert klientovi vrátane zbierky ID 190485 pre workshop Skriptor

Označte túto zbierku a dvakrát na ňu kliknite. V okne Zbierky (*Collections*) sa otvorí nové okno.

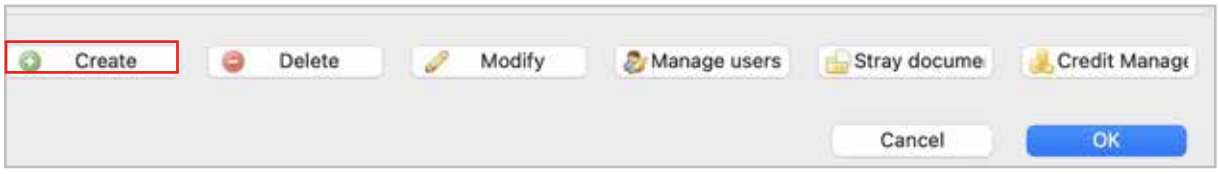


Obrázok 14 Práca so zbierkou

V tomto okne sa nachádzajú voľby pre prácu so zbierkami. Môžete mať viac zbierok a v každej zbierke viac dokumentov.

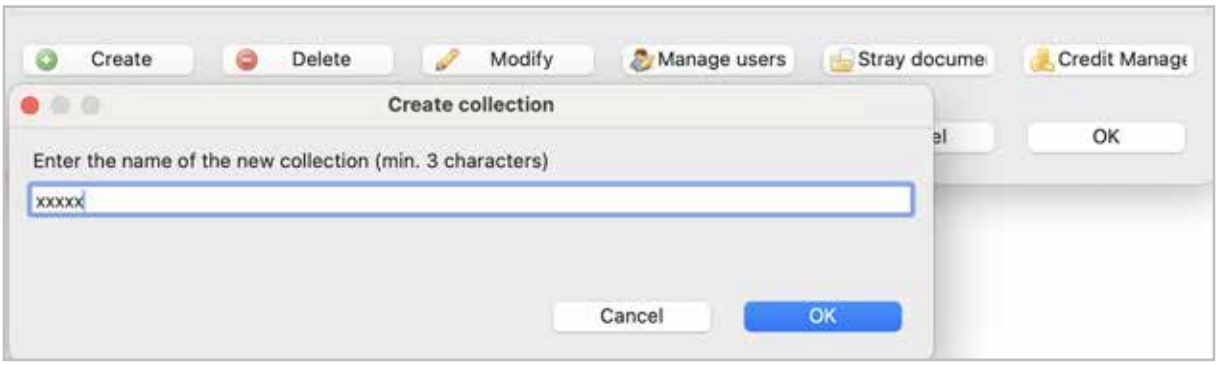
1. **Vytvoriť zbierku** (*Create*)
2. **Vymazať zbierku** (*Delete*)
3. **Upraviť zbierku** (*Modify*)
4. **Spravovať používateľov** (*Manage users*)
5. **Opustené dokumenty** (*Stray documents*)
6. **Správa kreditov** (*Credit Manager*)

2.3 Vytvoriť zbierku (*Create*)



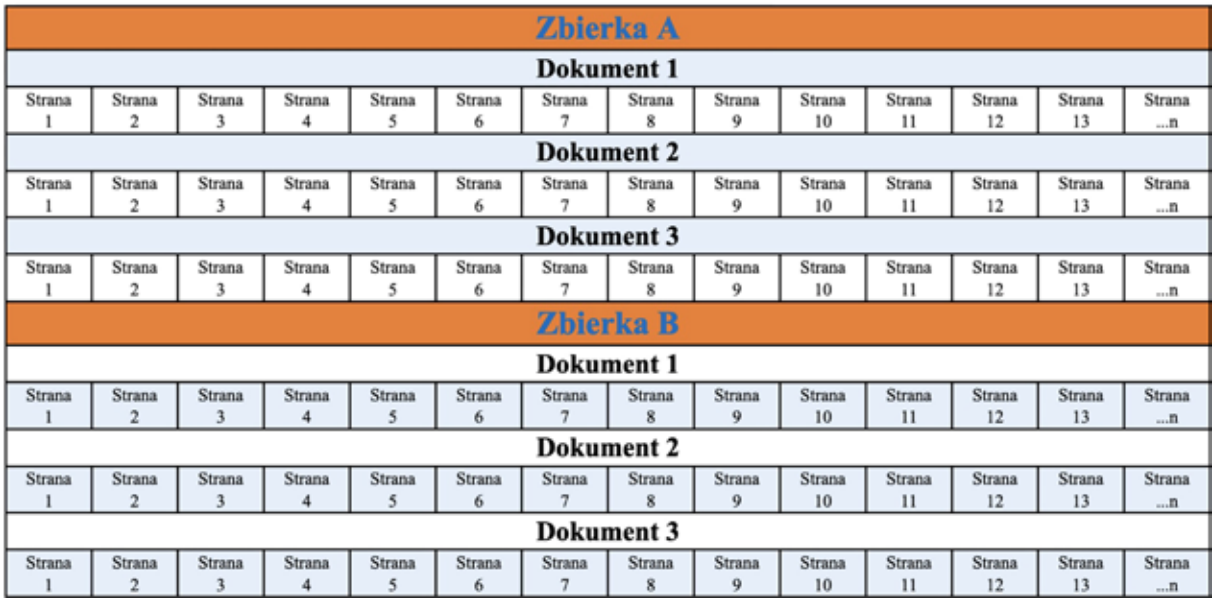
Obrázok 15 Tlačidlo voľby Create (Vytvoriť zbierku)

Do okna napíšte názov zbierky. Názov má mať minimálne 3 znaky. Stlačte tlačidlo OK.



Obrázok 16 Okno na pomenovanie novej zbierky

Dôležité je pochopiť, ako sú zbierky a dokumenty v Transkribe štruktúrované. Na obrázku nižšie vidieť logickú štruktúru zbierok Transkribu.



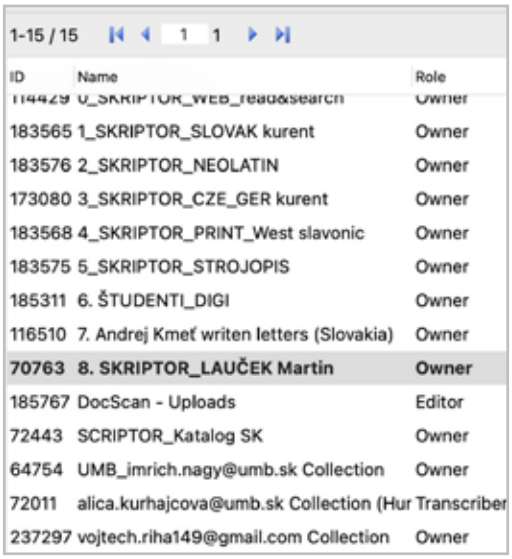
Obrázok 17 Štruktúra zbierok na platforme Transkribus

Dokumenty sú usporiadané v *zbierkach*.

Zbierku možno chápať ako *priečinok* obsahujúci dokumenty. Zbierky sa zvyčajne používajú na projektovom základe. Napríklad všetky dokumenty patriace do jedného projektu sú usporiadané v jednej zbierke. Príklad: zbierku pre účely projektu *Transkripcia korešpondencie J. M. Hurbana* môžeme nazvať *Hurban_listy*.

V jednej zbierke môže byť viac *dokumentov.* Dokumenty pozostávajú z jednej alebo viacerých strán dokumentu. Napríklad v zbierke *Hurban_listy* sú ako dokumenty jednotlivé listy. V projekte Skriptor má každý riešiteľ vytvorený vlastný účet – vlastný projekt a vlastné zbierky.

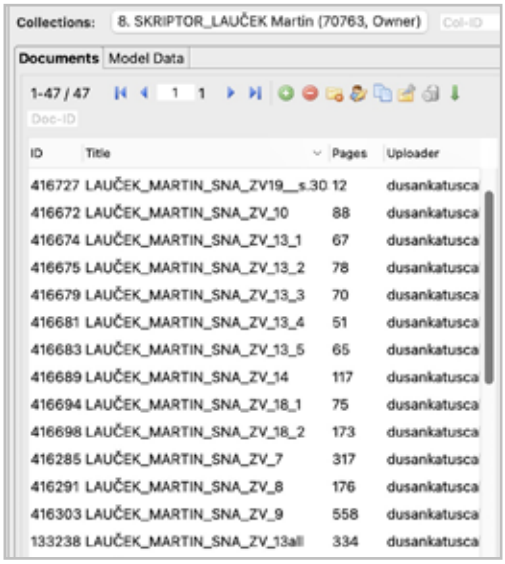
Sprístupnenie všetkých zbierok a dokumentov z projektu APVV Skriptor na internete je možné cez nástroj platformy *Read&search*. Všetky zbierky a dokumenty určené na zverejnenie na internete preto musia byť uložené v jednej zbierke (napr. 0_SKRIPTOR_WEB_read&search).



Obrázok 18 Zbierka ID 114429 určená na zverejnenie zbierok a dokumentov z projektu Skriptor cez stránku Read&search

Zbierky môžete vytvárať tak, aby zodpovedali organizácii fondov, zbierok a dokumentov v inštitúcii. Napríklad v archíve SNM v Martine je zbierka rukopisnej korešpondencie Andreja Kmeťa deponovaná v piatich škatuliach.

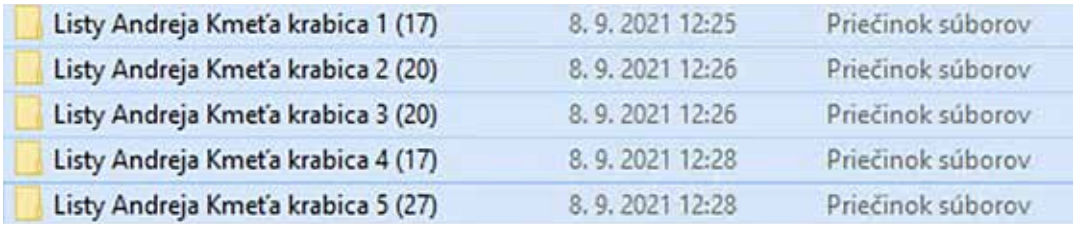
Dokumenty môže vlastník zbierky (*owner*) usporiadať tak, že uloží všetky listy Andreja Kmeťa do zbierky *Andrej Kmeť written letters*. V nej má napríklad dokument *List Kmeťa adresátke Balkovej* a 3 strany.



Obrázok 19 Príklad zbierky s dokumentmi a stranami

Ak chcete presne dodržať spôsob, akým sú dokumenty uložené v archíve, môžete pre každú z piatich škatúľ vytvoriť samostatnú zbierku s dokumentmi v tejto škatuli. Takto môžete mať napríklad vo svojej zložke päť zbierok a v každej zbierke dokumenty, listy podľa uloženia v škatuliach.

Poznámka: To, ako chcete mať usporiadané zbierky, by ste mali mať premyslené už pri snímaní dokumentov (skenovaní, fotografovaní).



Obrázok 20 Usporiadanie zbierok v osobnom počítači podľa škatúľ v archíve

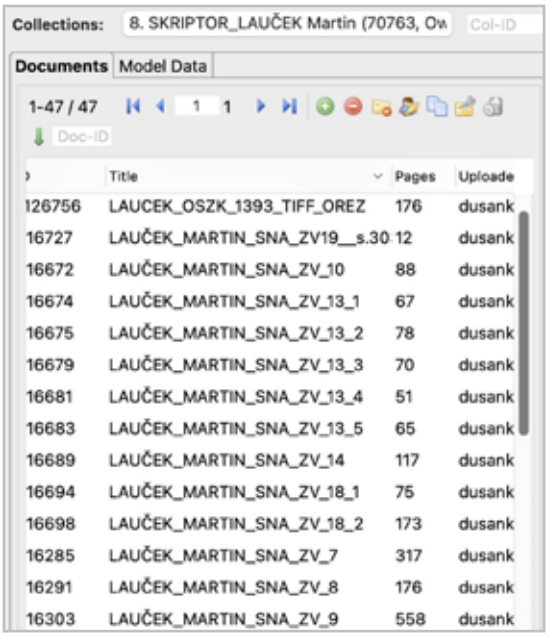
Ak chcete postupovať takto, vytvorte najprv vo svojej zložke päť zbierok a pomenujte ich. Napríklad *Kmeť 1*, *Kmeť 2*, *Kmeť 3*, *Kmeť 4*, *Kmeť 5*.

Do takto vytvorených zbierok nahrajte jednotlivé dokumenty (do *Kmeť 1* dokumenty zo škatule č. 1, atď.)

Ak chcete dodržať spôsob uloženia archívnych dokumentov (päť škatúľ), vytvorte 5 zbierok. V škatuli č. 1 je 17 zložiek (obalov s listami podľa adresátov). V zbierke *Listy Andreja Kmeťa* škatuľa č. 1 nahrajte všetky dokumenty (listy) podľa adresátov. Vytvárať 17 zbierok pre každú zložku by nemalo praktický zmysel. V prípade korešpondencie Andreja Kmeťa ide o homogénny fond: 5 škatúľ s listami usporiadanými podľa adresátov.

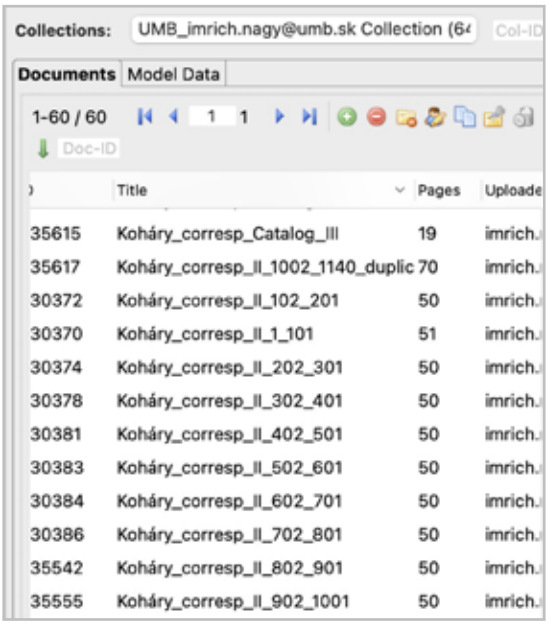
Všetky listy si už v osobnom počítači vopred pripravte ako súbory vo formáte PDF na import do platformy bez ohľadu na uloženie v archíve. Usporiadanie všetkých listov podľa adresátov v abecednom poradí je pre používateľa prijateľnejšie.

Alternatívne je možné do zbierky nahrať samostatne jednotlivé fyzické zväzky. Napríklad v zbierke Martina Laučeka *Collectanea* je základné rozdelenie podľa označovania zväzkov z archívov podľa miesta uloženia: SNK, z archívu SNM a z OsZK a potom podľa označenia v jednotlivých archívoch. Niektoré rozsiahlejšie zväzky sú rozdelené na menšie dokumenty kvôli experimentom v projekte, čo však nie je potrebné. Napríklad zväzok 13 je rozdelený na 5 častí.



Obrázok 21 Usporiadanie zväzkov podľa miesta uloženia

Niekedy je na účely experimentovania vhodné rozdeliť jednotlivé zväzky podľa počtu strán (napríklad 50 s.)



Obrázok 22 Usporiadanie rozdelených zväzkov v zbierke Kohary

Pomenovanie zbierok a dokumentov pripravte vopred už v procese snímania (skenovania). Digitálne dokumenty označujte spôsobom, ktorý je určený pre archívnu prax. Inštitúcie, ktoré sa rozhodnú pre transkripciu pomocou platformy Transkribus, môžu pomenovať zbierky a dokumenty tak, ako ich majú vo svojich fondoch.

2.3.1 Kontrola kvality pred importom

Kontrola kvality pred nahrávaním dokumentov do expert klienta je mimoriadne dôležitá, pretože umožňuje udržiavať poriadok v platforme, organizovať dokumenty a pripravovať dokumenty na editovanie a sprístupnenie na internete cez nástroj *Read&search*.

Skontrolujte:

1. úplnosť dokumentu,
2. kvalitu snímania (ostrosť, kontrast, farebnosť, úplnosť snímanej plochy – strany, presvity),
3. orientáciu strán,
4. poradie strán,
5. formát snímania.

Po snímaní je z dôvodu archivácie, zálohovania a ďalšej manipulácie potrebné vytvoriť:

1) *archívnu kópiu* v úložisku alebo na externých nosičoch. Ide o adresár, do ktorého uložíte obrázky v „surovom“, needitovanom formáte, v akom boli nasnímané (fotografované, skenované) JPG/TIFF/RAW/PNG. Zodpovedný: systémový administrátor inštitúcie – informatik,
2) *derivovanú kópiu*. Adresár, v ktorom budú už „surové“ obrázky po postprocesingu, teda po následnej úprave a kontrole kvality upravené, opravené, orezané, úplné so správnou orientáciou a v dobrej kvalite. Kópiu tohto adresára v najkvalitnejšom dohodnutom formáte nahrajte na nosič (CD, SD, USB, externý disk) alebo na digitálne úložisko a:
   a) poskytnite ho inštitúcii ako vlastníkovi alebo správcovi zbierky,
   b) uložte ho v inštitúcii na účely neverejného prístupu nedostupného cez internet. Určte miesto uloženia a zodpovednú osobu (systémový knihovník, kurátor a pod.).
3) *pracovnú kópiu* v adresári na svojom počítači s dokumentmi v derivovanom fomáte PDF, z ktorého nahráte zbierky a dokumenty na platformu Transkribus. Zodpovedný: manažér projektu transkripcie,
4) *transkribovanú kópiu* v adresári alebo adresároch s exportovanými súbormi, ktoré sú výsledkom transkripcie. Zodpovedný: manažér projektu transkripcie,
5) *datasety* so súbormi *Ground Truth* a vlastnými modelmi. Zodpovedný: manažér projektu transkripcie a systémový administrátor.

Dokumenty nahrávate (importujete, uploadujete) na platformu Transkribus preto, aby ste ich automaticky transkribovali a sprístupnili odbornej a širšej verejnosti.

Cez expert klienta platformy Transkribus má zmysel transkribovať väčšie fondy, zbierky a dokumenty, teda stovky až tisícky strán. Menšie zbierky a dokumenty je možné zatiaľ transkribovať prostredníctvom Transkribus Lite.

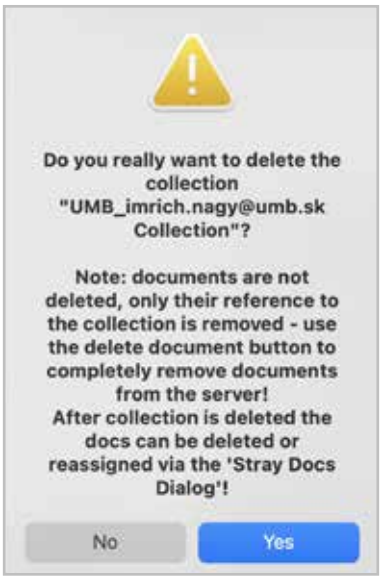
V expert klientovi platformy Transkribus pracujete s tými dokumentmi, pre transkripciu ktorých je potrebné vytvoriť vlastné modely transkripcie, pretože dostupné modely sú neuspokojivé.

2.4 Vymazať zbierku (*Delete*)

Obrázok 23 Tlačidlo voľby Delete (Vymazať zbierku)

Kliknite na tlačidlo Vymazať zbierku (*Delete*), ak chcete zbierku vymazať. Zobrazí sa varovanie, či naozaj chcete vymazať označenú zbierku.

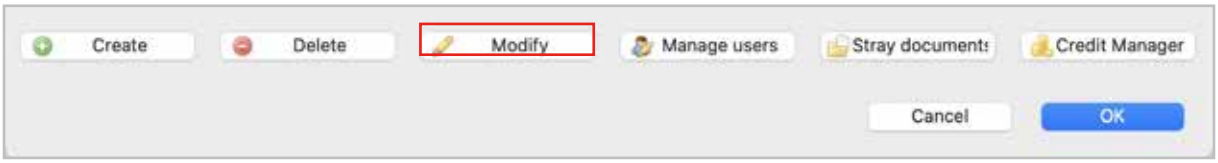
Poznámka: Dokumenty v zbierke sa fyzicky nevymažú, odstráni sa iba ich prepojenie na zbierku.



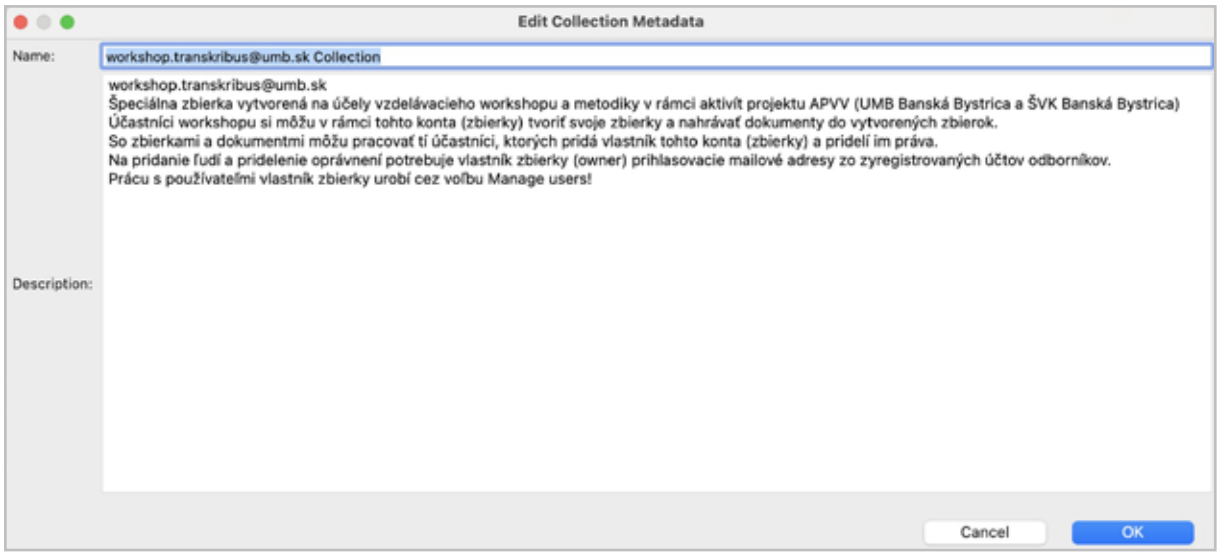
Obrázok 24 Upozornenie pred vymazaním zbierky

Po odstránení zbierky je možné dokumenty vymazať alebo zmeniť priradenie prostredníctvom funkcie *Stray Docs Dialog*.

2.5 Upraviť zierku (*Modify*)



Obrázok 25 Tlačidlo voľby Modify (Upraviť zbierku)



Obrázok 26 Okno na popis zbierky a metadáta

Do okna s popisom zbierky napíšte text vzťahujúci sa na zbierku: popis zbierky, metadáta a pod. Tieto údaje môžete kedykoľvek zmeniť.

2.6 Spravovať používateľov (*Manage users*)

Obrázok 27 Tlačidlo voľby Manage users (Spravovať používateľov)

Nového používateľa/používateľov je možné pridať cez voľbu Spravovanie používateľov (*Manage users*). Táto funkcia umožňuje:

- zapísať e-mail nového používateľa zbierky cez voľbu *Username/E-Mail*,
- vyhľadať nového používateľa *v* účtoch Transkribus cez voľbu *Find users*,
- určiť oprávnenia na prácu so zbierkou pre nového používateľa pomocou voľby *Change Role.*

Do okna vpravo dolu E-mail používateľa (*Username/E-Mail*) napíšte e-mailovú adresu účtu používateľa na platforme Transkribus (z prihlasovacích údajov).

Kliknite na ikonu Nájsť používateľov (*Find users*). Ak existuje, nové meno používateľa sa zobrazí v okne Používateľské meno (*Username/Name*).

Kliknite na meno používateľa. Vyberte voľbu Pridať používateľa (*Add user*).

Kliknutím na meno existujúceho používateľa v ľavom okne sa aktivuje možnosť Zmeniť rolu používateľa (*Change Role*) na platforme Transkribus:

1. vlastník (O*wner*) – najvyššie oprávnenia, môže robiť všetky zmeny a úpravy,
2. editor (*Editor*) – môže editovať text,
3. transkriber (*Transcriber*) – môže transkribovať/prepisovať text,
4. čitateľ (*Reader*) – môže text iba čítať, t. j. má len pasívny prístup k zbierke.

Voľba Pridať do inej zbierky (*Add to other collection*) umožňuje pridať označeného používateľa do inej zbierky cez dialógové okno.

Obrázok 28 Pridanie nového používateľa do zbierky a určenie práv

2.7 Opustené dokumenty (*Stray documents*)

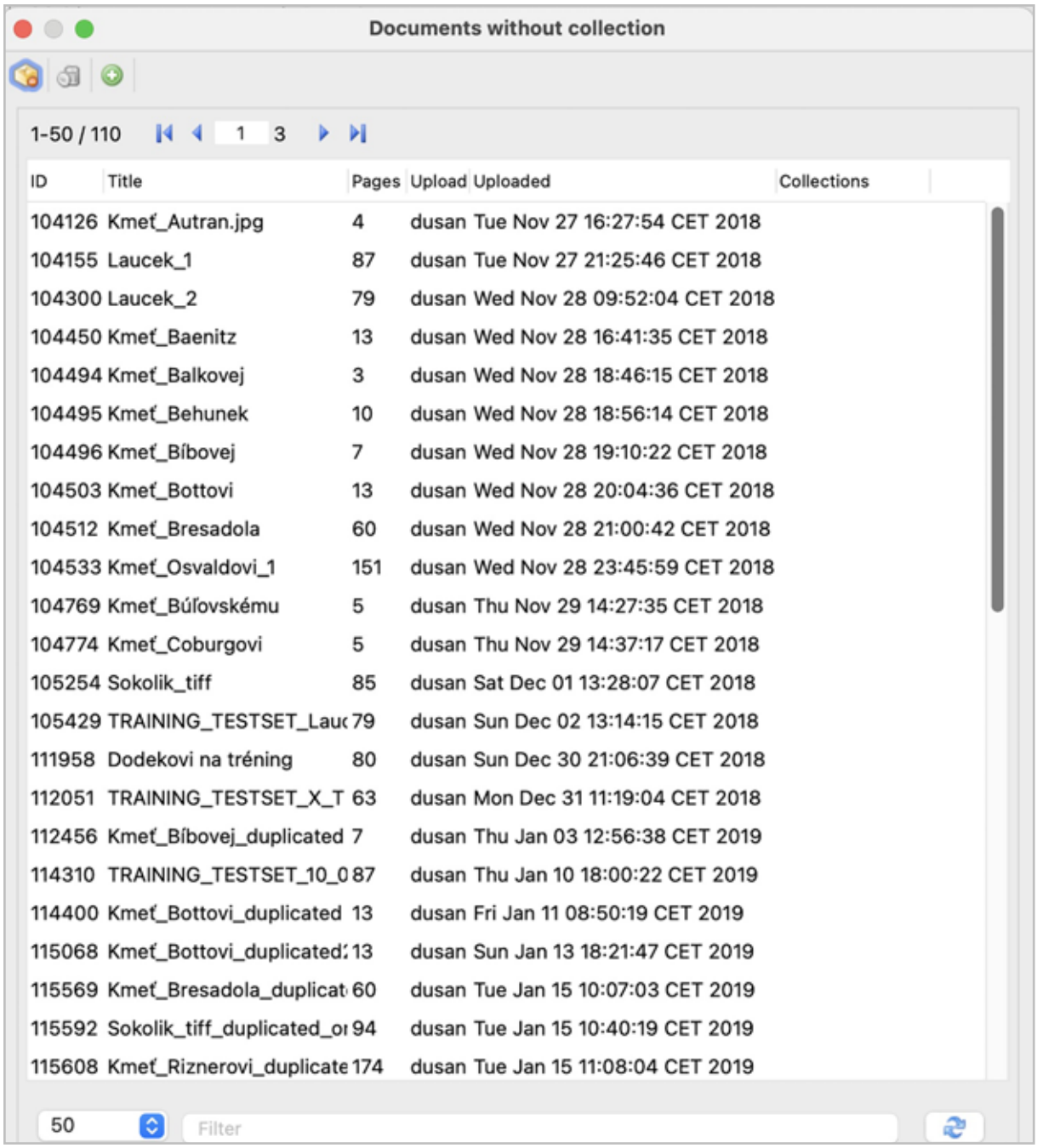
Obrázok 29 Tlačidlo voľby Stray documents (Opustené dokumenty)

Voľba Opustené dokumenty (*Stray documents*) súvisí s funkciou Vymazať zbierku (*Delete*). Kliknutím na voľbu Vymazať zbierku sa vymaže názov zbierky, ale dokumenty z vymazanej zbierky sú stále uložené na serveri.

Kliknutím na voľbu Opustené dokumenty (*Stray documents*) sa zobrazí zoznam dokumentov, ktoré nepatria do žiadnej zbierky.

Vpravo hore v okne Opustené dokumenty (*Stray documents*) sú ďalšie voľby, ktoré umožňujú:

a) vymazať označený dokument,
b) vymazať všetky dokumenty,
c) presunúť označený dokument do inej zbierky.



Obrázok 30 Dokumenty, ktoré boli vymazané a momentálne nepatria do žiadnej zbierky

3 Príprava dokumentu na automatickú transkripciu

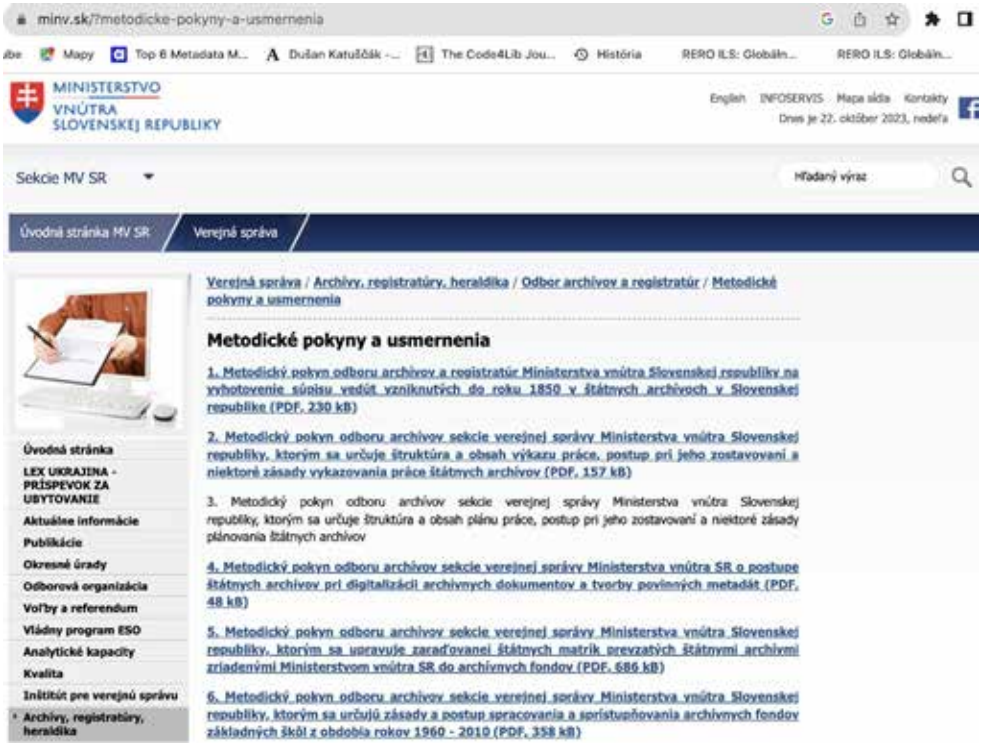
Prípravná fáza na prácu s dokumentom na platforme Transkribus zahŕňa kritériá, ktoré by mali digitalizáty spĺňať, ich správny popis, samotné vyhotovenie kvalitných digitalizátov a následný import do expert klienta.

3.1 Kritériá výberu digitalizátov

V podmienkach slovenskej archívnej praxe upravuje výber archívnych dokumentov určených na digitalizáciu Metodický pokyn odboru archívov sekcie verejnej správy Ministerstva vnútra SR o postupe štátnych archívov pri digitalizácii archívnych dokumentov a tvorby povinných metadát č. SVS-OA-2011/23406-001 z roku 2011. Pokyn definuje prednostný výber archívnych fondov a archívnych zbierok ako aj technické parametre systematickej digitalizácie. Odporúča na digitalizáciu dokumenty sprístupnené archívnou pomôckou, chronologicky spred roka 1526, často využívané bádateľmi a fyzicky najohrozenejšie. Z hľadiska technických parametrov stanovuje vyhotovenie digitálnych kópií z originálnych dokumentov alebo mikrofilmov. Každý záber digitálnej kópie archívneho dokumentu má byť uložený ako samostatný súbor s popisným reťazcom pozostávajúcim zo šiestich, resp. siedmich častí v tvare:

SK_aaaa_fffff_iiiii_ssss_x.ext

Štruktúru reťazca tvorí kód krajiny (SK), štvormiestne číslo archívu (z číselníka štátnych archívov SR, v príklade aaaa), päťmiestne číslo archívneho súboru (z aplikačného programového vybavenia AFondy, v príklade fffff), päťmiestne označenie inventárneho čísla alebo signatúry archívneho dokumentu (iiiii), štvormiestne poradové číslo snímky v rámci inventárneho čísla alebo signatúry (ssss), znak x označuje písmeno alebo číslicu rozlišujúce digitálnu kópiu (konzervačnú, pre interné potreby, alebo na študijné účely), prípona .ext označuje grafický formát (JPEG alebo TIFF).

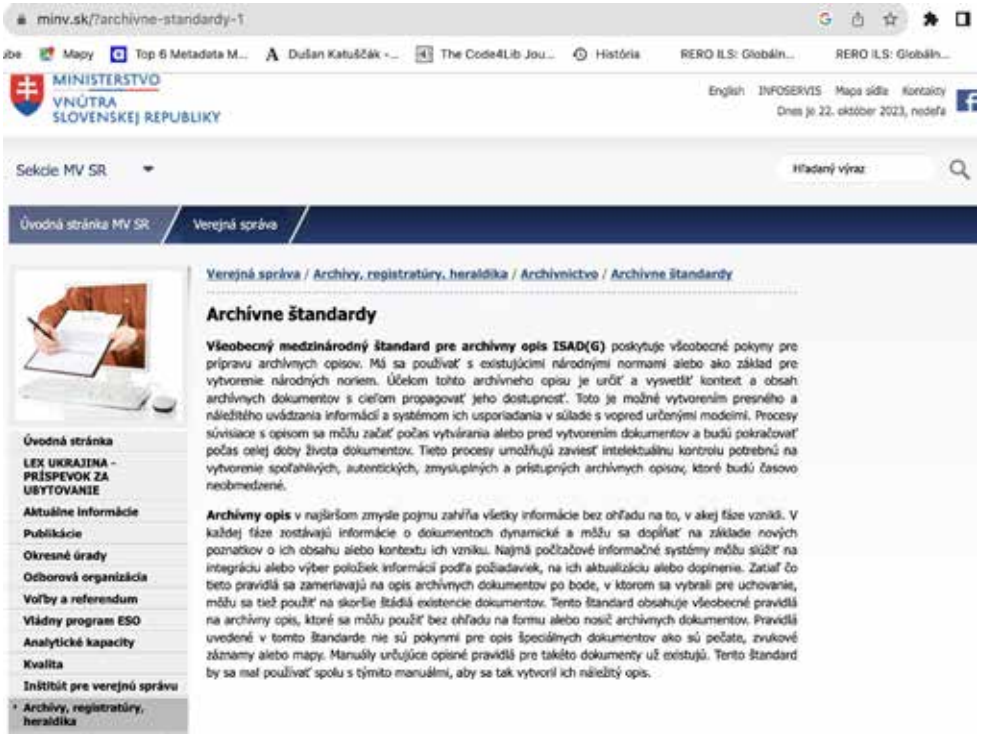


Obrázok 31 Metodický pokyn MV SR č. SVS-OA-2011/23406-001

Voľba dokumentu pre platformu Transkribus sa oproti tomu vyznačuje niektorými špecifikami. Na rozdiel od bežnej digitalizácie v pamäťových inštitúciách nie sú prioritou poškodené a ďalšou manipuláciou ohrozené archívne dokumenty. Vzhľadom na ich stav zachovania (porušenosť, fragmentárnosť) nie sú príliš vhodné pre segmentáciu, transkripciu a nadväzujúce postupy. Naopak vhodnejšie sú intaktne zachované archívne dokumenty. Na prácu v Transkribe sú najefektívnejšie rozsiahle rukopisy vyhotovené jednou pisárskou rukou najlepšie v krátkom časovom úseku. Z hľadiska potrieb bádateľov a vedeckého výskumu sem možno zaradiť napríklad matričnú agendu, kanonické vizitácie, sčítacie operáty, parcelné protokoly a pod. S prihliadnutím na charakter platformy je vhodné vyberať dokumenty odrážajúce špecifiká slovenského kultúrneho okruhu, ktoré sú atraktívne aj pre zahraničných užívateľov.

3.2 Popis fondov, zbierok a dokumentov

Význam presného popisu je osobitne dôležitý pre digitálne objekty nadobúdajúce podobu elektronického informačného zdroja. Na rozdiel od fyzického vyhotovenia vznikajú digitálne dokumenty iba vďaka softvéru. Stráca sa tým jedinečnosť a proveniencia fyzicky zachovaných archívnych dokumentov. Elektronické dokumenty preto nadobúdajú zvýšené požiadavky na overovanie faktov (*fact-checking*) s dôrazom na dôveryhodnosť, spoľahlivosť, ale aj pôvodnú provenienciu a hierarchiu. Popis a citovanie má umožňovať dohľadanie fyzicky zachovaných zdrojov. Vzhľadom na medzinárodný obsah a rozšírenie platformy Transkribus je pre digitalizované dokumenty (materiál textovej povahy) vhodné využívať štandardizované medzinárodné popisy a normy.



Obrázok 32 Všeobecný medzinárodný štandard pre popis archívnej jednotky

Medzinárodná rada archívov schválila a zverejnila niekoľko štandardov zjednocujúcich popis archívnych dokumentov, vrátane digitalizátov. Všeobecný medzinárodný štandard pre popis archívnej jednotky (*General International Standard for Archival Description – ISAD(G)*) vychádza z provenienčného princípu a definuje dvadsaťšesť položiek popisu. Odbor archívov a registratúr Ministerstva vnútra SR sprístupnil slovenský preklad druhého vydania štandardu z roku 1999 aj

s príkladmi viacúrovňových popisov pre sieť štátnych archívov na Slovensku. Keďže archívne dokumenty uchovávajú aj iné subjekty, napr. kultúrne inštitúcie a súkromní vlastníci, vznikli ďalšie normy. Medzinárodný štandard pre archívne autoritné záznamy právnických osôb, fyzických osôb a rodín (*International Standard of Archival Authority Record for Corporate Bodies, Persons and Families – ISAAR(CPF)*) je rozšírený najmä v mimoeurópskom priestore. Ďalšie úpravy obsahuje Všeobecný medzinárodný štandard pre popis inštitúcií s archívnymi dokumentmi (*International Standard for Describing Institutions with Archival Holdings – ISDIAH*).

Možnosti pre jednotný popis digitalizovaných dokumentov zo štátnych archívov, ale aj cirkevných archívov, knižníc, múzeí, galérií, pamiatkových úradov, vedeckých ústavov a pod. poskytuje štruktúra popisného reťazca pre platformu Transkribus – projekt *Skriptor*. Na rozdiel od archívnej terminológie metodika vychádza z určenia pre digitálny repozitár. Obsahuje fixné názvy zbierok a súborov určených na automatickú transkripciu s dôrazom na prehľadnosť a zrozumiteľnosť. Štruktúru reťazca tvorí názov zbierky (kolekcie), názov podzbierky (subkolekcie) a zdroj/vlastník. Za fixnými časťami nasledujú premenlivé hodnoty dopĺňané podľa konkrétnej situácie a podľa skenovaných objektov.

Tieto hodnoty sú najmä:

1) označenie (číslo) zväzku (signatúra),
2) počet listov,
3) rok(y) RRRR alebo RRRR-RRRR.

Celý názov entity určený na nahratie do Transkribu môže mať napríklad takúto štruktúru:

LAUČEK_MARTIN_SNA_ZV_13_5
Skriptor_Hurban listy_SNKLA_2A3_1875_Pauliny-Tóth Viliam
Visitatio canonica_CV18_DABB

Ak je snímaný objekt určený pre digitálny repozitár, vloží sa na začiatok reťazca referenčný kód a názov jednotky popisu – v prípade štátnych archívov podľa ISAD(G) kód krajiny, archívu, fondu alebo zbierky.

3.3 ScanTent a DocScan pre archívy a knižnice

DocScan a *ScanTent* sú nové nástroje, ktoré pomáhajú snímať historické dokumenty na účely transkripcie v dobrej kvalite. Informácie o nástrojoch sú dostupné z hlavnej stránky READ-COOP na https://readcoop.eu/transkribus/?sc=Transkribus.

V bádateľniach archívov bádatelia používajú na snímanie vlastné zariadenia, fotoaparáty, mobilné telefóny, tablety a podobne. ScanTent a DocScan sú prijateľnou alternatívou k bežným zariadeniam na snímanie dokumentov v archívoch a knižniciach.

ScanTent a DocScan je výborným riešením pre inštitúcie, ktoré nemajú pre používateľov k dispozícii kvalitnejšie stolové skenery alebo ktoré ešte nemajú zdigitalizované svoje dokumenty prístupné pre používateľov.

Obrázky snímané týmto spôsobom je možné poskytnúť inštitúcii na dohodnutom nosiči alebo na uloženie do inštitucionálneho digitálneho repozitára archívu alebo knižnice. Ak však máte možnosť rozhodnúť sa medzi zariadeniami ScanTent a DocScan a profesionálnym skenerom dokumentov, uprednostnite profesionálny skener.

Pre digitalizáciu platí zásada, že snímanie – skenovanie sa robí v najvyššej možnej kvalite, na najvyššej dosiahnuteľnej úrovni. Kvalita snímaných obrázkov je kľúčová pre efektívnu tran-

skripciu. Skúsenosti ukazujú, že kvalita snímania by mala byť okolo 600 DPI. Historické rukopisy predstavujú *de facto* špecifickú grafiku, pre ktorú sa niekedy odporúča snímanie v kvalite 900 až 1200 DPI. Práca s vysokokvalitnými obrázkami však môže vyžadovať postprocesing, čiže následné spracovanie pomocou špeciálnych softvérov na úpravu obrazu.

Pri práci s DocScan a ScanTent trvá naskenovanie knihy, teda fyzického zväzku s 300 stranami, približne 12 –15 minút. To je 150 obrázkov, pretože v zariadení sa snímajú naraz obidve strany otvoreného zväzku prakticky až do veľkosti A3. Spravidla teda budete môcť nasnímať **viac ako 500 obrázkov za hodinu.**

3.3.1 ScanTent

ScanTent je možné získať zakúpením priamo z hlavnej stránky po voľbe ScanTent.

Je optimálnym riešením na snímanie voľných alebo zviazaných dokumentov v bádateľniach – pre nízkonákladové a vysokokvalitné snímanie (skenovanie). Cena ScanTentu je aktuálne 239,00 € vrátane 20% DPH plus poštovné.

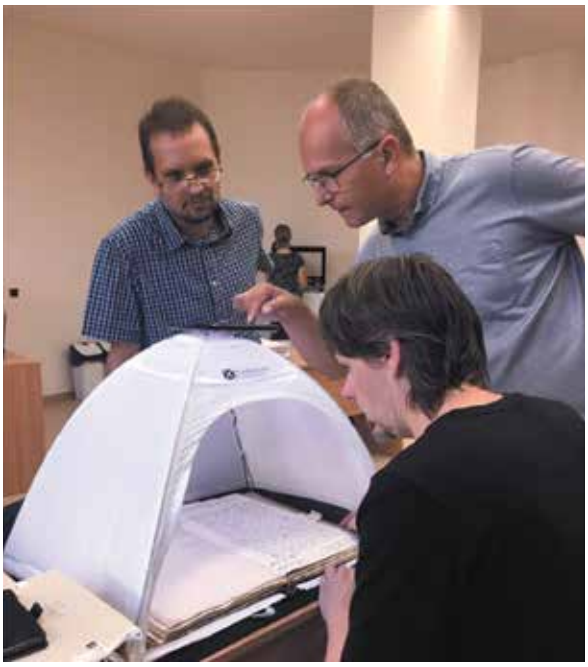
Niektoré inštitúcie majú pre bádateľov a čitateľov v študovniach a bádateľniach desiatky zariadení ScanTent. Napríklad Francúzska národná knižnica ich mala 40 v roku 2023. Ak chcete získať ďalšie informácie, kontaktujte scantent@caa.tuwien.ac.at

V katalógu tovarov a služieb je to na účely obstarania a evidencie majetku tovar v rámci skupiny „statívy na fotoaparáty“ pod číslom 90 391.

Hlavné funkcie ScanTentu:

- **profesionálne fotografické prostredie** na snímanie vysokokvalitných obrázkov bez dodatočného svetla. Stan je z nylonovej hodvábnej látky s vnútornými príchytkami na led osvetlenie,
- **LED osvetlenie s USB napájaním** pre nepriame osvetlenie dokumentov – pripojenie na notebook alebo iný zdroj (napríklad powerbank),
- **tmavá plstená látka** na základni ako optimálny podklad,
- **veľká základná plocha**, takže používatelia môžu vložiť ruky do zariadenia a držať zviazané dokumenty otvorené oboma rukami,
- skenovanie dokumentov **veľkosti približne A3 alebo aj o niečo väčších**,
- ľahký (500 gramov) a **skladateľný**, zmestí sa do malého puzdra.

Obrázok 33 Prototyp ScanTent použitý na snímanie zväzkov Martina Laučeka v SNA v Bratislave a v SNM v Martine (2018)



Obrázok 34 Novší model ScanTent použitý na snímanie v Diecéznom archíve Banskobystrického biskupstva v Badíne v rámci projektu Skriptor (10.09.2020)

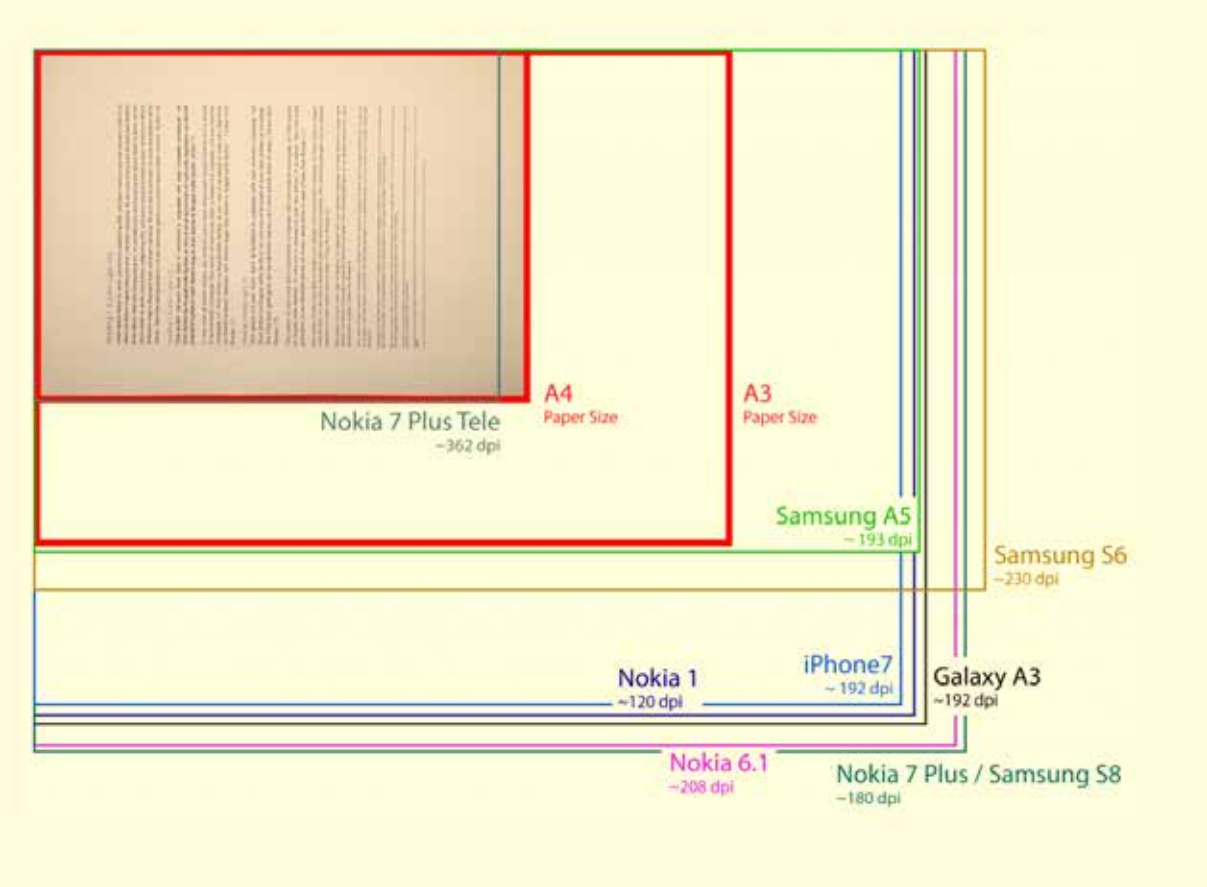
Popis častí ScanTentu:

1. kompatibilita s každým smartfónom (*Compatible with every smartphone*)
2. nepriame LED osvetlenie (*Indirect LED Lighting*)
3. protišmyková plocha pod smartfón (*Anti-slip phone-mount*)
4. difúzne osvetlenie (*Diffusion of ambient light*)
5. nylonová hodvábna látka (*Nylon silk fabric*)
6. voľne dostupná aplikácia DocScan s umelou inteligenciou (*Free AI-Powered DocScan App*)
7. veľká základná plocha (A3 alebo časopisecký formát tabloid 280 mm × 430 mm) (*Large base area*)

Obrázok 35 Komponenty ScanTentu

Rýchly prehľad montáže ScanTentu nájdete vo videu https://youtu.be/iL2WNNi5VEI

Po nastavení zariadenia ScanTent môžete začať so snímaním – fotografovaním. Ak okolité svetlo nie je dostatočné (čo sa stáva veľmi zriedka), zapnite svetlá do prenosného počítača alebo USB zásuvky.

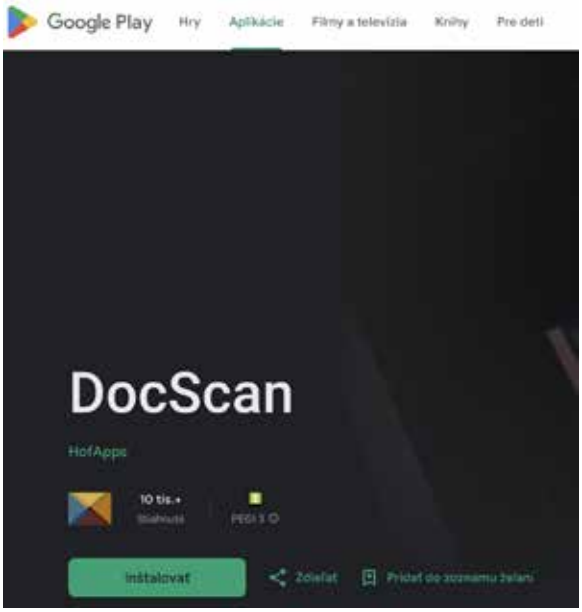


Obrázok 36 Experimenty vývojového tímu s rôznymi smartfónmi

Vývojový tím platformy Transkribus testoval osem rôznych smartfónov a meral ich rozlíšenie. Zorné pole a rozlíšenie jednotlivých smartfónov môžete vidieť na obrázku vyššie. DocScan nepodporuje telefóny Nokia 7 Plus Tele a iPhone 7.

3.3.2 Aplikácia DocScan

Aplikácia DocScan sa používa so zariadením ScanTent a je to softvér, ktorý vyvinula Technická univerzita vo Viedni v rámci európskeho projektu READ. Aplikáciu DocScan si môžete bezplatne stiahnuť z obchodu Google Play na https://play.google.com/store/apps/details?id=at.ac.tuwien.caa.docscan



Obrázok 37 Stránka Google Play na stiahnutie a inštaláciu DocScan

Aplikácia DocScan bola vyvinutá špeciálne na digitalizáciu kníh a archívnych dokumentov pomocou smartfónu. V súčasnosti je k dispozícii prednostne pre telefóny so systémom Android.

DocScan je určený na skenovanie historických dokumentov v kombinácii so ScanTentom. Zobrazuje strany v živom náhľade a robí skeny v dostatočnej kvalite pre platformu Transkribus. V automatickom režime *Series* sníma obrázok po otočení stránky po pripojení k zariadeniu ScanTent. Umožňuje teda rýchlo skenovať knihy alebo dokumenty bez interakcie s vaším mobilom.

Hlavné funkcie DocScan:

- rýchla a spoľahlivá detekcia stránok dokumentu,
- jednoduchý režim (*Single*) na manuálne snímanie jednotlivých obrázkov,
- sériový režim (*Series*) na automatické snímanie obrázkov (pohyb je detekovaný automaticky). Po otočení automaticky sníma ďalší obraz dvojstrany,
- schopnosť otáčať a orezávať stránky,
- priame nahrávanie dokumentov na server Transkribus.

Výhody:

- vysoká kvalita obrazu – moderné inteligentné telefóny poskytujú vynikajúcu kvalitu obrazu s vysokým rozlíšením,
- nákladovo efektívne – ako pre koncového používateľa, tak aj pre knižnicu/archív,
- žiadne licenčné poplatky,
- nie je potrebná žiadna používateľská podpora z archívu alebo knižnice – používatelia sa s aplikáciou DocScan rýchlo zoznámia sami,
- priateľské k autorským právam – používatelia snímajú a ukladajú obrázky na svojom vlastnom zariadení, nie na tých, ktoré vlastní knižnica alebo archív,
- DocScan ponúka možnosť „masového skenovania“, kde je možné obrázky vytvorené používateľmi pridať do digitálnych fondov knižnice alebo archívu.

3.3.3 Bezpečnosť údajov v aplikácii DocScan

Bezpečnosť sa začína pochopením toho, ako vývojový tím zhromažďuje a zdieľa vaše údaje. Postupy ochrany osobných údajov a zabezpečenia sa môžu líšiť v závislosti od vášho používania, regiónu a veku. Vývojový tím aktuálne poskytuje nasledujúce informácie a môže ich časom aktualizovať.

- Táto aplikácia môže zdieľať tieto typy údajov s tretími stranami: miesto a osobné údaje.
- Táto aplikácia môže zhromažďovať tieto typy údajov: osobné informácie, fotografie a videá, súbory a dokumenty.
- Dáta sú pri prenose šifrované.
- Údaje nie je možné vymazať.

3.3.4 Snímanie pomocou ScanTent a DocScan

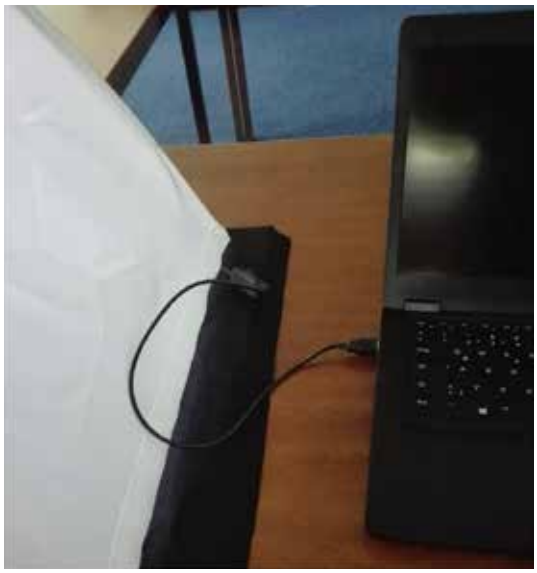
Položte smartfón na podložku v hornej časti ScanTent tak, aby šošovka fotoaparátu smerovala nadol. Šošovka by mala byť zarovnaná s otvorom v hornej časti zariadenia.

Polohu smartfónu je vhodné nastaviť paralelne vzhľadom na snímanú plochu a orientáciu strany. Poloha smartfónu by mala zostať počas snímania dokumentu v stabilnej a rovnakej polohe vo vzťahu k snímanému dokumentu, aby dodatočne nebolo potrebné korigovať orientáciu strán alebo opakovane snímať nesprávne snímanú plochu strany. Smer snímania ukazuje obrázok písmena „T“.

Dôležité: ScanTent umiestnite vyššie alebo nižšie podľa toho, či chcete pri snímaní sedieť alebo stáť.

Displej smartfónu musí byť rovnobežný so smerom dokumentu. Ak stojíte pred ScanTentom, musíte vidieť na displej a vedieť čítať správy DocScanu na smartfóne. Mobil by mal byť orientovaný rovnako ako strana.

Obrazovku DocScan je možné pre pohodlie snímania zrkadliť na ďalšom počítači, takže DocScan môžete vidieť a ovládať cez počítač, nielen cez smartfón položený na ScanTente.

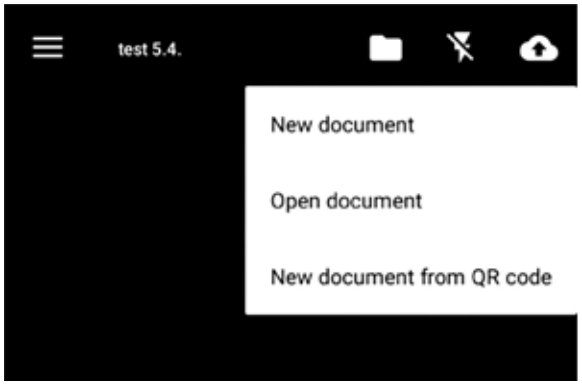


Obrázok 38 Pripojenie osvetlenia LED k notebooku

Obrázok 39 ScanTent pripravený na snímanie smartfónom

3.3.5 Práca s aplikáciou DocScan

Otvorte aplikáciu kliknutím na ikonu DocScan v telefóne.

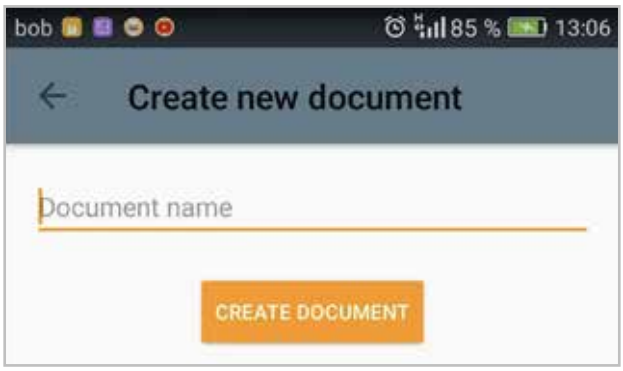


Obrázok 40 Spojenie DocScan s aplikáciou Transkribus za účelom prenosu údajov z DocScanu cez prihlásenie „burger“

Obrázok 41 Plocha aplikácie DocScan prihláseného používateľa

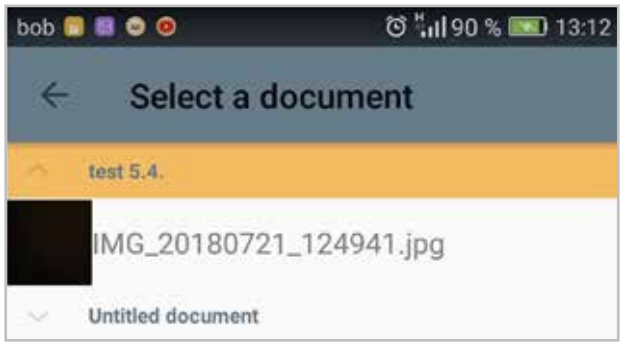
Kliknite na voľbu Dokumenty (*Documents*). Priraďte svojmu dokumentu názov.

Vyberte možnosť Vytvoriť dokument (*Create document*). Všetky obrázky, ktoré následne nasnímate, budú uložené pod týmto menom vo vašom telefóne a zostanú v ňom, aj keď ich nahráte do Transkribu.



Obrázok 42 Vytvoriť a popísať nový snímaný dokument

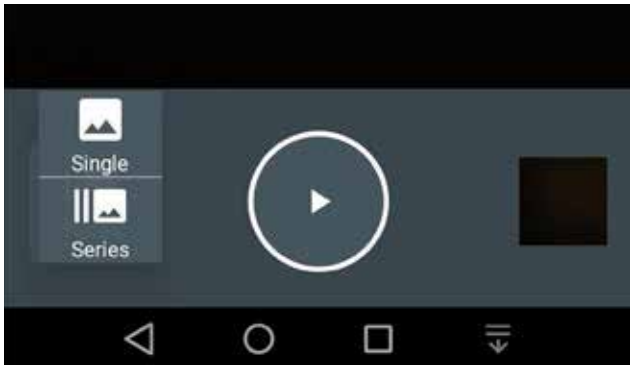
K existujúcim dokumentom môžete pridať nový dokument kliknutím na ikonu „+“. Kliknite na ikonu priečinka v pravej hornej časti aplikácie. Potom vyberte možnosť Otvoriť dokument. Vyberte názov existujúceho dokumentu a zvoľte možnosť Použiť vybratý dokument.



Obrázok 43 Pridanie nového dokumentu k existujúcim dokumentom

Po popísaní dokumentu môžete začať skenovať. Umiestnite telefón na vrchnú časť zariadenia ScanTent.

Na hlavnej stránke kliknite na možnosť *Camera.* Môžete si vybrať, či chcete nasnímať jednotlivé obrázky manuálne alebo nastaviť aplikáciu tak, aby automaticky zachytávala obrázok pri každom otočení stránky. Môžete si vybrať z možností *Manual*/*Single* alebo *Automatic*/*Series* v ľavej dolnej časti aplikácie.



Obrázok 44 Režimy snímania: Manual/Single režim, Automatic/Series režim

Snímanie spustite kliknutím na ikonu fotoaparátu v krúžku.

Na telefóne zapnite zvuk. Ten upozorní na otočenie strany. Otočenie a zosnímanie indikuje aj svetelný signál, ak ho máte zapnutý.

Strany otáčajte opatrne, neponáhľajte sa, aby DocScan stačil zaostriť, a aby správne snímal celú plochu. Unáhlené pohyby môžu spôsobiť nedostatočné zaostrenie a rozmazanie snímaného obrazu.

Po snímaní dokumentu je potrebná kontrola kvality snímania alebo postprocesing, čiže následné spracovanie obrazov v dokumente. Zamerajte sa na úplnosť, možné duplicity, orientáciu strán a pod.

Proces snímania je možné vrátiť cez ikonu troch vodorovných čiarok, tzv. „burger“. Ku kamere sa dostanete cez tú istú ikonu.

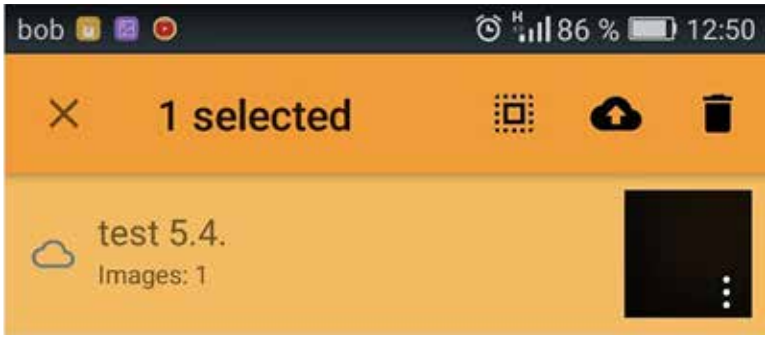
3.3.5.1 Odoslanie dokumentu na platformu Transkribus

Obrázok 45 Nahrávanie do Transkribu cez ikonu Cloud

Stlačte ikonu cloudu v pravej hornej časti aplikácie. V prípade potreby sa prihláste do svojho účtu Transkribus.

Vyberte dokument, ktorý chcete nahrať do Transkribu. Ešte raz stlačte ikonu cloudu.

Otvorte Transkribus na svojom počítači. Svoje nahraté dokumenty nájdete v zbierke s názvom *DocScan – Uploads*.



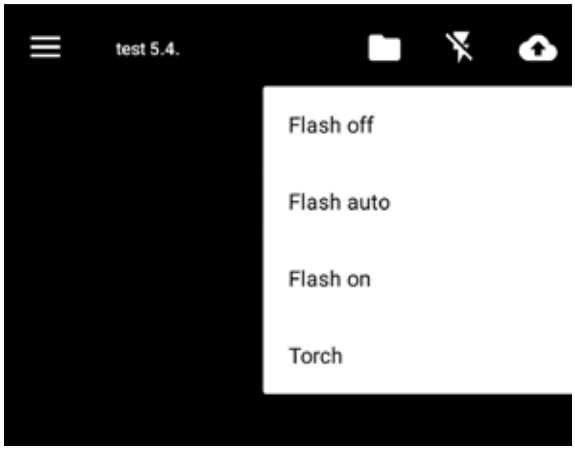
Obrázok 46 Výber súboru na nahratie do Transkribu

Nahrávanie do platformy Transkribus je zvyčajne pripravené za niekoľko minút. Ak odovzdávate veľké množstvo snímok, môže to trvať o niečo dlhšie.

3.3.5.2 Nastavenia

Ďalšie nastavenia nájdete a upravíte kliknutím na ikonu „burger” vľavo hore a výberom možnosti Nastavenia.

Blesk nastavíte stlačením ikony blesku v pravom hornom rohu aplikácie. Na výber sú štyri možnosti: vypnutý blesk (*Flash off*), automatický blesk (*Flash auto*), zapnutý blesk (*Flash on*) a svetlo (*Torch*).

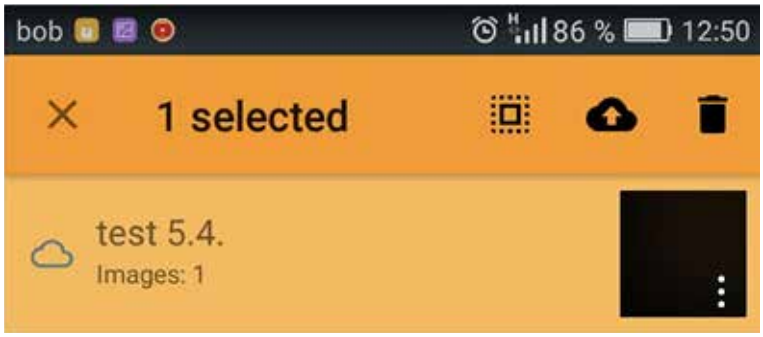


Obrázok 47 Nastavenie blesku

3.3.5.3 Automatické orezávanie, otáčanie a mazanie

Na orezanie a otočenie obrázkov podľa potreby môžete použiť DocScan.

1. Po nasnímaní obrázka stlačením miniatúry v pravom dolnom rohu aplikácie otvorte nastavenia úprav.

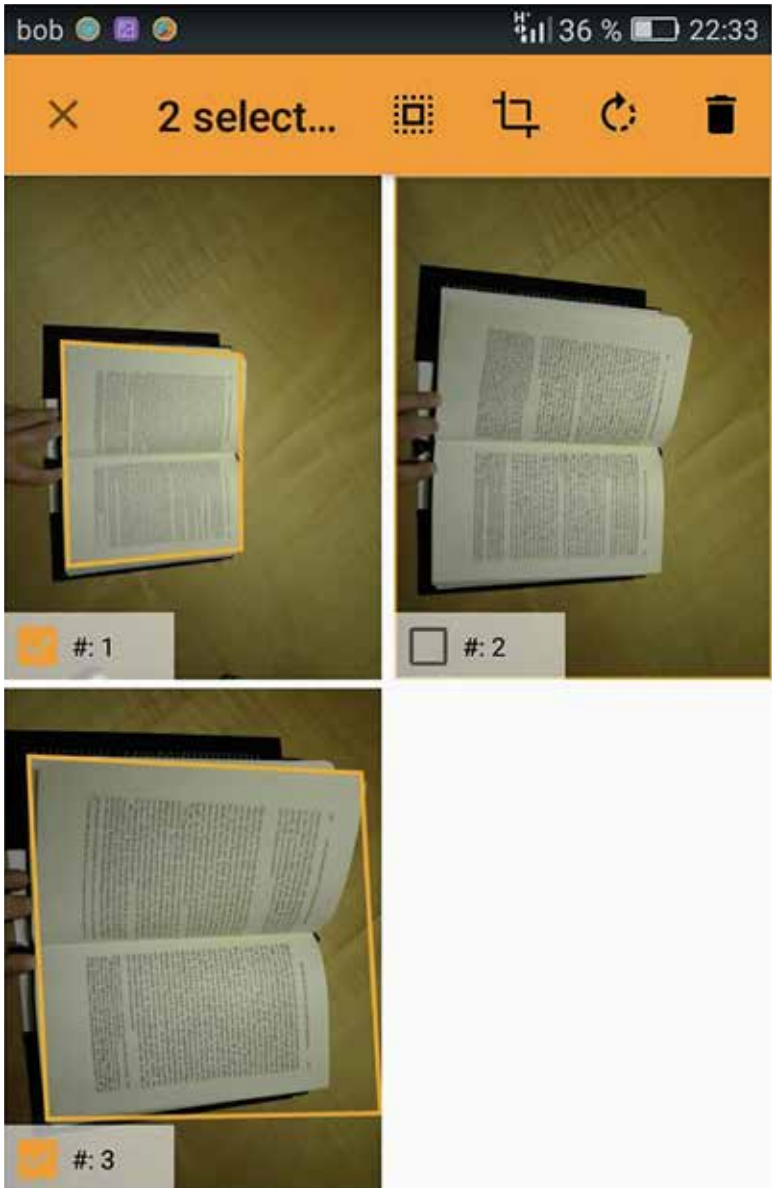


Obrázok 48 Výber strán dokumentu na orezanie cez miniatúru

2. Všetky strany sa zobrazia v žltých rámoch.

Poznámka: Keď je aktivované orezanie, do žltého rámu sa pridá niekoľko pixelov, takže na obrázku sa zobrazí celá strana.

3. Označte súbory, ktoré chcete orezať.



Obrázok 49 Výber strán na orezanie

Vďaka funkcii automatického orezania nemusíte presúvať rámy do správnej polohy, aplikácia to za vás urobí automaticky.

3.3.5.4 Manuálne orezanie

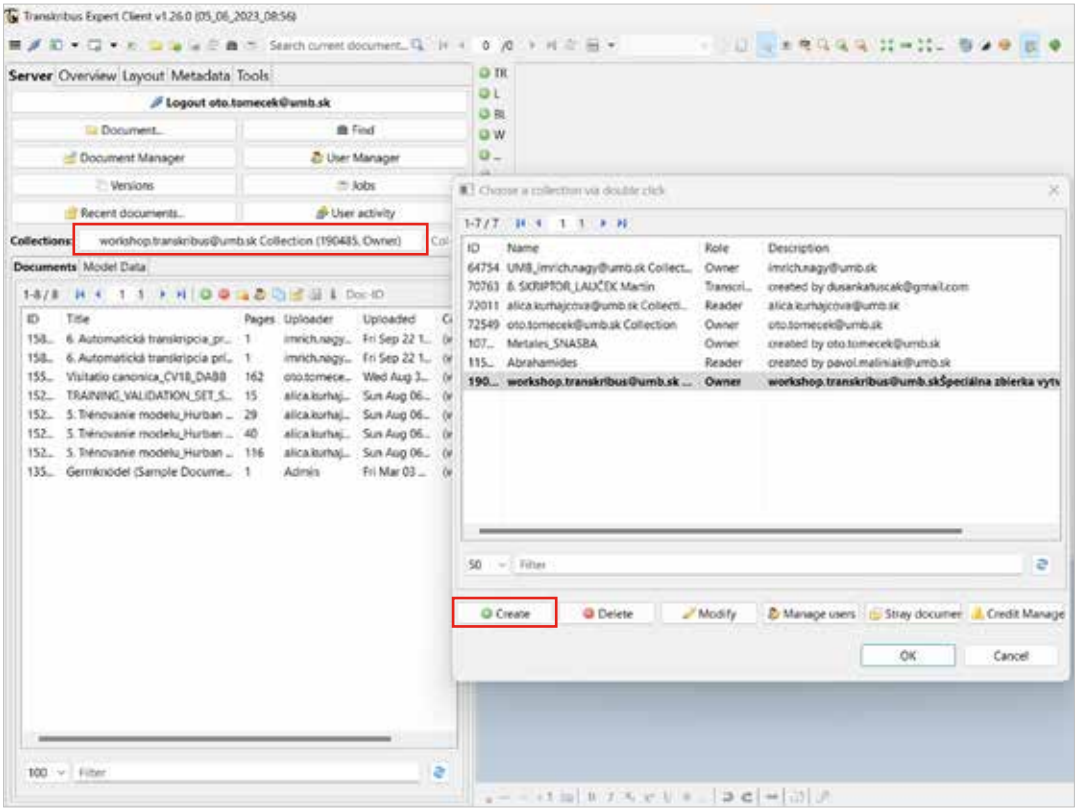
1. Kliknite na ikonu orezania v spodnej časti obrazovky.
2. Potiahnite rohy obrázka do požadovanej polohy.
3. Kliknite na ikonu orezania v pravom hornom rohu obrazovky a uložte orezaný obrázok.
4. Na ďalšej obrazovke kliknite na ikonu uloženia.

Obrázky môžete otáčať, zdieľať alebo odstrániť (zahodiť do koša) tak, že vyberiete potrebné strany a kliknete na príslušnú ikonu.

Obrázok 50 Voľby operácií otočiť, zdieľať, orezať, zahodiť do koša

3.4 Importovanie digitalizátov do Transkribu

Pred začatím importovania digitalizovaných dokumentov (digitalizátov) na server platformy Transkribus expert klient si najprv zvoľte zbierku, do ktorej chcete digitalizáty importovať. Zvoľte možnosť Zbierky (*Collections*). Následne po otvorení príslušného okna vyberte spomedzi existujúcich zbierok, alebo si vytvorte vlastnú zbierku (*Create*). Do takto zvolenej zbierky budete následne importovať pripravené digitalizáty dokumentu určeného na neskoršie transkribovanie.



Obrázok 51 Výber existujúcej zbierky alebo vytvorenie novej zbierky

Po kliknutí na príslušnú zbierku vyberte z hlavného menu voľbu Importovať dokument (*Import Document(s)*).

Do Transkribu je možné naimportovať dokumenty priamo prostredníctvom aplikácie DocScan (pozri kapitolu 3.3.2 Aplikácia DocScan), ďalej stiahnutím z internetu (napríklad dostupné digitalizované dokumenty zo stránok pamäťových inštitúcií) alebo ako samostatné vopred pripravené (naskenované alebo nafotené) dokumenty.

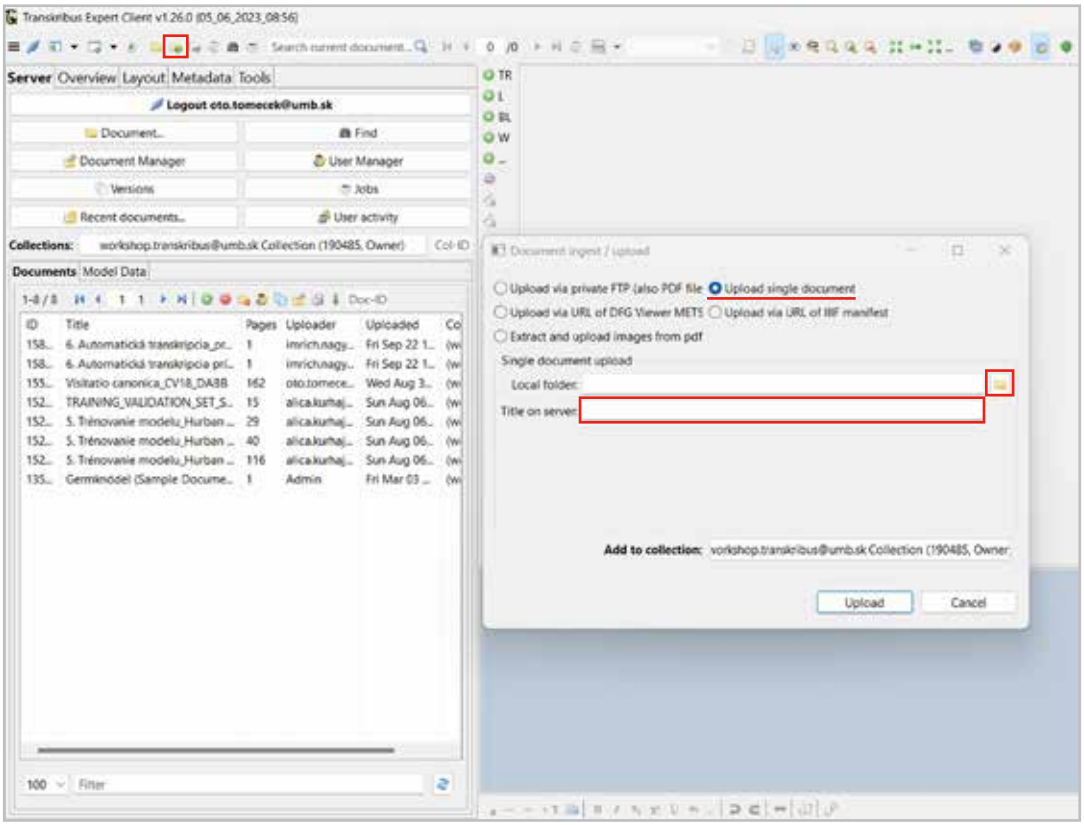
Importovanie digitalizátov a následná práca v expert klientovi je možná len so súbormi vo formátoch PDF, JPEG, PNG a TIFF. Obrazové súbory by mali byť pred importom do Transkribu pripravené v osobitnom priečinku vášho počítača.

Po kliknutí na voľbu *Import Document(s)* sa otvorí nové okno s ponukou možných spôsobov nahratia digitalizátov na platformy Transkribus. Zvolený dokument môžete vložiť piatimi rôznymi spôsobmi, a to označením (zakliknutím) jednej z nasledovných možností:

1) *Upload via private FTP*
2) *Upload via URL of DFG Viewer METS*
3) *Extract and upload images from pdf*
4) *Upload single document*
5) *Upload via URL of IIIF manifest*

Po zvolení preferovanej voľby sa zmení vizuál príslušného okna. Následne je potrebné doplniť požadované údaje. Pri prvej voľbe *Upload via private FTP* vyberáte dokument priamo z prostredia platformy Transkribus. Pri druhej a piatej voľbe, teda *Upload via URL of DFG Viewer METS* alebo *Upload via URL of IIIF manifest*, je potrebné do príslušného okna vložiť URL adresu stránky, kde sa nachádzajú vybrané digitalizáty. Pri tretej voľbe *Extract and upload images from pdf* je potrebné vybrať z lokálneho priečinka (*Folder*) formátu PDF.

Najjednoduchšou možnosťou v prípade nasnímania dokumentu do viacerých JPEG súborov je zvolenie štvrtého spôsobu, teda vloženie samostatného dokumentu. Po zvolení uvedenej voľby, ktorá je v prostredí Transkribu prednastavená, vyberte príslušný priečinok (*Local folder*) s pripravenými dokumentami a pomenujte ho vlastným názvom (*Title on server*). Podľa tohto názvu budú importované dokumenty na serveri neskôr ľahko identifikovateľné. Po týchto krokoch môžete začať s procesom nahrávania dokumentov na server potvrdením tlačidla *Upload* v spodnej časti okna.

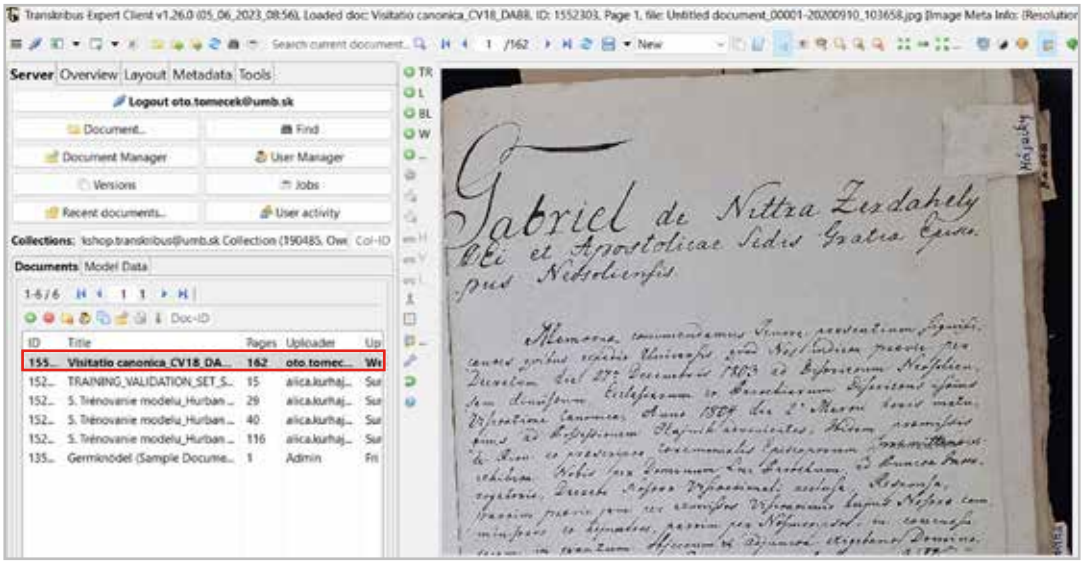


Obrázok 52 Možnosti importovania digitalizátov

Prenos všetkých dokumentov môže byť zdĺhavejší. Dĺžka importovania dokumentov na server závisí od aktuálnej vyťaženosti samotného serveru, predovšetkým však od veľkosti prenáša-

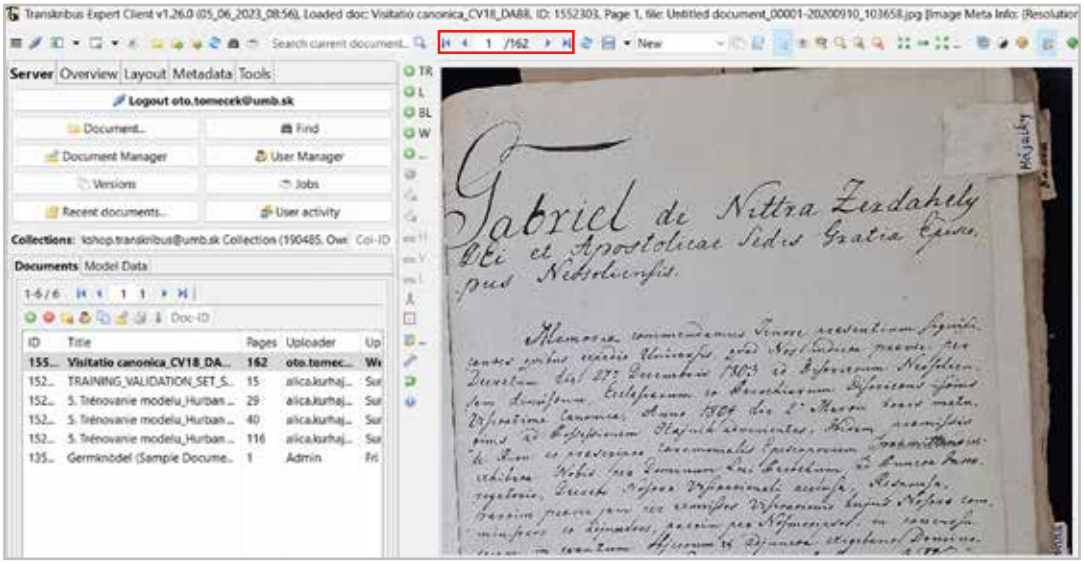
ných dokumentov (digitalizátov). Pri nahrávaní dokumentov na server vo veľkosti väčšej ako 1 GB je potrebné počítať s tým, že dĺžka procesu môže trvať aj viac ako jednu hodinu. Úmerne s narastajúcim množstvom GB sa násobí aj časová hodnota prenosu dokumentov na server.

Po ukončení importovania digitalizátov sa na obrazovke objaví okno oznamujúce ukončenie procesu nahrávania dokumentov. Aby sa importované dokumenty stali viditeľnými, je potrebné odísť zo zvolenej zbierky prekliknutím na inú zbierku a následne sa do nej opätovne vrátiť. Oba uvedené kroky realizujte kliknutím na voľbu *Collections*. Po opätovnom zvolení príslušnej zbierky sa v ľavej časti obrazovky objavia všetky dokumenty, ktoré sú do zbierky zaradené. Prostredníctvom dvojkliku na príslušnú položku v zozname (*Title*) sa v pravej časti obrazovky otvorí úvodná strana nahratého dokumentu.



Obrázok 53 Otvorenie príslušného nahratého dokumentu

Medzi jednotlivými stranami dokumentu prechádzajte prostredníctvom tlačidiel na hornom ovládacom paneli, alebo manuálnym vpísaním čísla požadovaného digitalizátu do políčka označujúceho číslo digitalizátu a potvrdením uvedenej voľby prostredníctvom tlačidla ENTER na svojom počítači.



Obrázok 54 Preklikávanie medzi jednotlivými stranami importovaného dokumentu

4 Segmentácia dokumentov v Transkribe

Keď máte dokument nahratý v Transkribus expert klientovi, môžete začať s analýzou rozloženia (*Layout Analysis*). Výsledkom analýzy je segmentácia nasnímaných snímok dokumentu, t. j. identifikácia jednotlivých prvkov, rozlíšenie štruktúry, horizontálnej orientácie textu a určenie poradia čítania textu.

Pri segmentácii sa uplatňuje metóda analýzy obrazu a textovej analýzy, ktorých výsledkom je členenie textu na časti, resp. objekty. Tie sa následne prepájajú s textom, ktorý bude výsledkom transkripcie.

Každý objekt segmentácie určuje, kde sa nachádzajú:

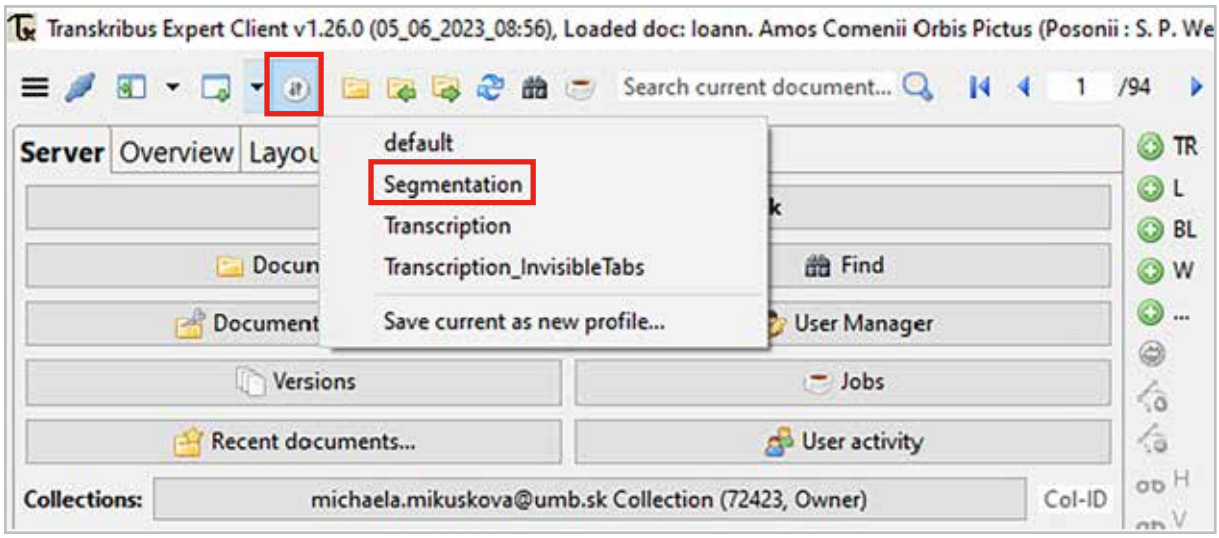
- **textové rámce** (*Text Regions, TR*) – vymedzujú oblasti s textom, môže ísť o hlavný text dokumentu, čísla strán, marginálie, tabuľky a i., označené sú zeleným rámom,
- **riadkové rámce** (*Line Regions, LR*) – vymedzujú riadky v rámci textových rámcov, označené sú tyrkysovým rámom,
- **základné čiary** (*Baselines, BL*) – vymedzujú čiaru, ktorá sa tiahne pozdĺž spodnej strany riadka. Ide o najdôležitejší referenčný bod na rozpoznávanie textu, na základe ktorého sa softvér učí čítať jednotlivé znaky. V závislosti od zvoleného profilu sú označené fialovou alebo červenou farbou,
- okrajové a nadbytočné časti dokumentu, ktoré nie sú dôležité pre proces transkripcie a trénovania modelu.

Súradnice objektov sa v procese segmentácie ukladajú do súboru príslušnej stránky dokumentu. Správna segmentácia textu výrazne ovplyvňuje prepis dokumentu, kvalitu vytrénovaného modelu, korekciu transkripcie a proces spracovania prepísaného textu.

Transkribus má k dispozícii niekoľko profilov zobrazenia dokumentu, ktoré sa zobrazia kliknutím na ikonu  (*Profiles*) v hlavnom menu:

- štandardné zobrazenie (*Default*) – na snímke dokumentu sú zobrazené všetky segmentované objekty, riadky sú označené tyrkysovým rámom, základné čiary sú zvýraznené fialovou farbou, pod snímkou dokumentu sa zobrazuje pole textového editora, v ktorom sa zapisuje/zobrazuje transkripcia dokumentu,
- **segmentácia** (*Segmentation*) – na snímke dokumentu sú zobrazené všetky základné čiary červenou farbou, snímka dokumentu je viditeľná na celej pravej strane expert klienta,
- **transkripcia** (*Transcription*) – pod snímkou dokumentu sa zobrazuje pole textového editora, v ktorom sa zapisuje/zobrazuje transkripcia dokumentu, na snímke dokumentu sa zobrazuje len riadok, s ktorým aktuálne pracujete.

Na analýzu rozloženia odporúčame použiť profil Segmentácia (*Segmentation*), pretože v ňom sú najlepšie viditeľné chyby, ktoré vznikli v procese segmentácie. Kvalitnú analýzu môžete zrealizovať aj v profile Štandardné zobrazenie (*Default*).



Obrázok 55 Výber profilov segmentácie a transkripcie

4.1 Spôsoby segmentácie

Analýzu rozloženia (*Layout analysis*) môžete urobiť dvomi spôsobmi:

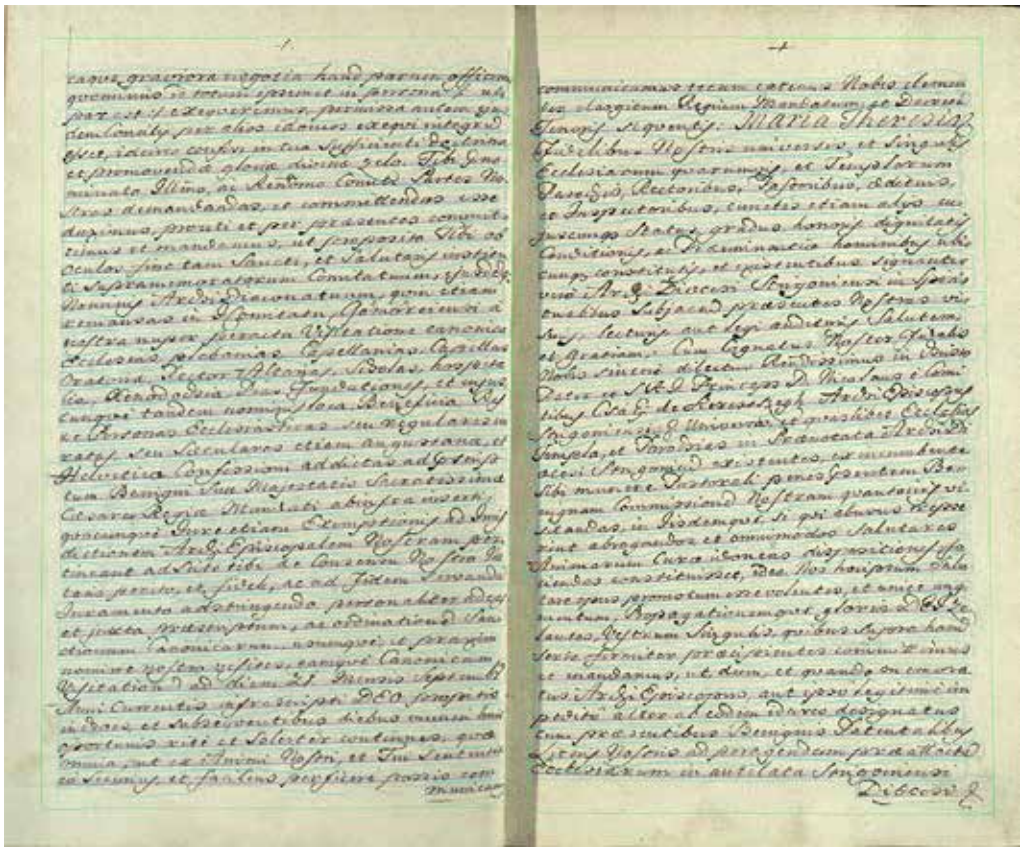
- **automaticky** – označenie textových rámcov, oblastí riadkov a základných čiar necháte urobiť výlučne softvér,
- **manuálne** – spočíva v manuálnom vytvorení textových rámcov a automatickej segmentácii riadkových rámcov a základných čiar.

Objekty segmentácie, textové rámce a základné čiary môžete označiť výlučne manuálnym spôsobom, t. j. bez použitia funkcií automatickej segmentácie. Ide však o veľmi prácny a časovo náročný proces. Nástroje na tvorbu objektov sú popísané v kapitole *4.2 Opravy po automatickej a manuálnej segmentácii*.

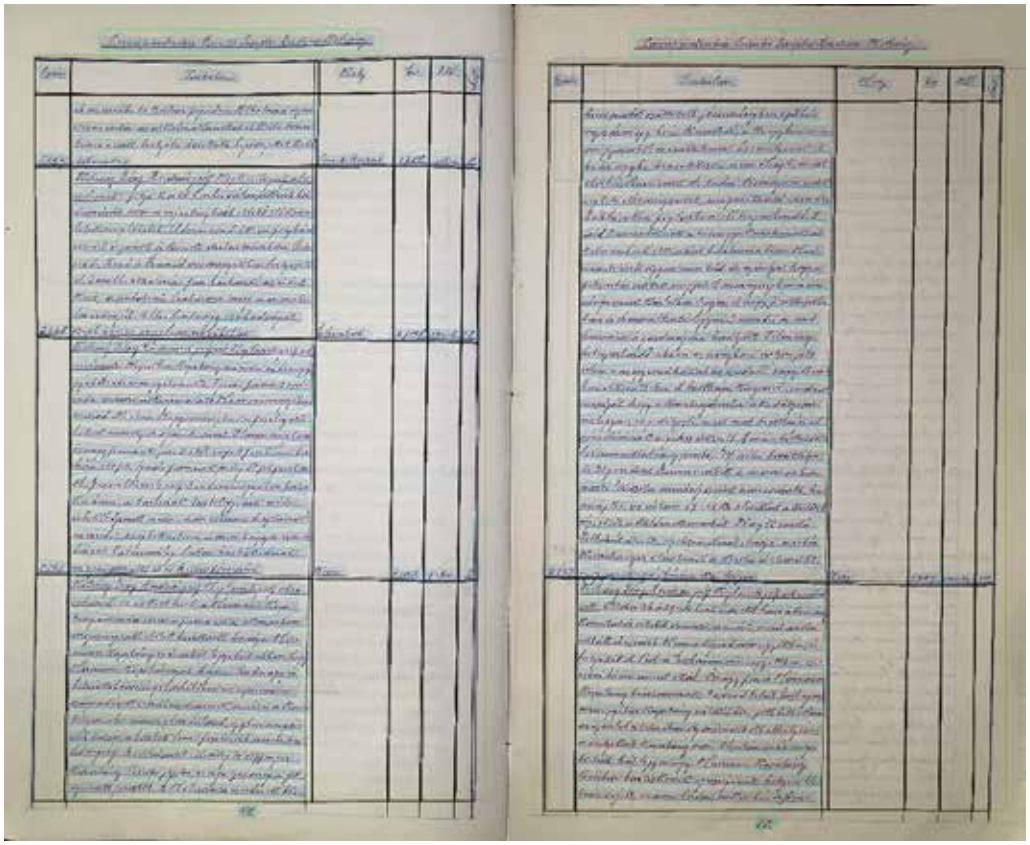
Výber spôsobu segmentácie závisí od štruktúry a obsahu dokumentu, s ktorým pracujete. Nesprávne zvolený typ segmentácie môže viesť k časovo náročným opravám. Automatická segmentácia rozpozná, kde sa text na snímke dokumentu graficky nachádza, rozpozná základné textové rámce a riadky v nich, ale nerozlišuje typ obsahu. Text vo vytvorených blokoch zoradí podľa súradníc objektov na snímke, spravidla od ľavého horného rohu smerom nadol. Automatickú segmentáciu je preto vhodné použiť na dokumenty s jednoduchou štruktúrou a jasným poradím riadkov.

Pri komplikovanom rozložení textu je však potrebné definovať viac blokov textu. Manuálnu segmentáciu je vhodné použiť pri členitom obsahu a zložitejšej štruktúre textu dokumentu, napr. ak text obsahuje poznámky pod čiarou, stĺpce, tabuľky, alebo sa v dokumente vyskytujú marginálie a i.

Na obrázkoch nižšie môžete vidieť príklady dokumentov vhodných na automatickú a manuálnu segmentáciu.



Obrázok 56 Príklad dokumentu s jasnou štruktúrou poradia blokov textu a riadkov vhodného na automatickú segmentáciu



Obrázok 57 Príklad dokumentu so zložitou štruktúrou textu vhodného na manuálnu segmentáciu

4.1.1 Automatická segmentácia

Pri automatickej segmentácii softvér na snímke dokumentu sám vyznačí textové rámce, oblasti riadkov a základné čiary.

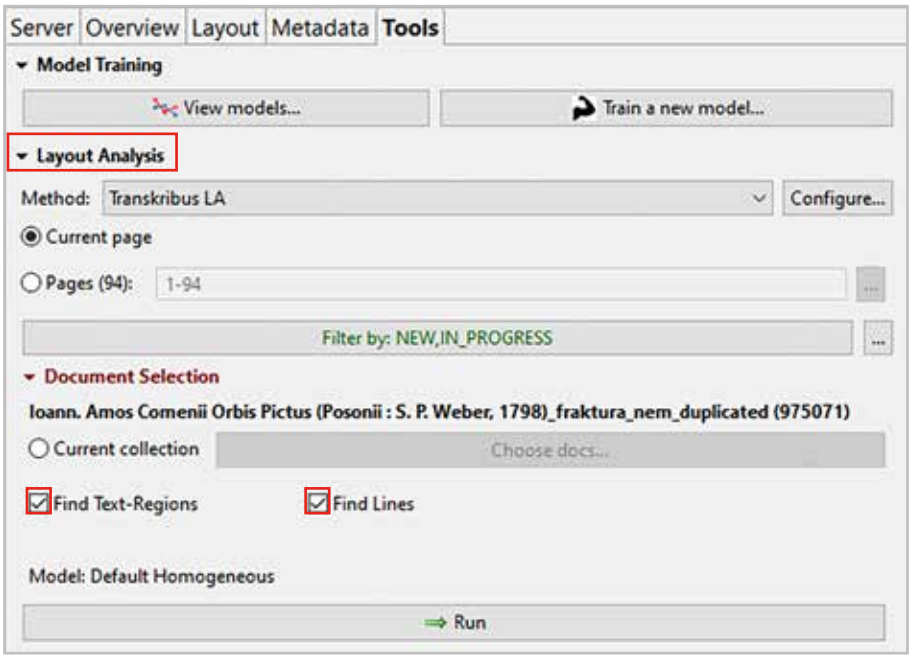
Nastavenie a spustenie automatickej segmentácie
Otvorte záložku **Tools** na ľavej strane klienta pod hlavným menu. Prejdite do sekcie ***Layout Analysis***. Pred spustením segmentácie treba:

1. nastaviť metódu (*Method*) – automaticky býva prednastavená *Transkribus LA*. Kliknutím na šípku na konci riadka sa otvoria ďalšie metódy (*Kraken*, *Printed Block Detection* a *Separator Detection*). Voľbu metódy vyberte podľa typu dokumentu, s ktorým pracujete. *Transkribus LA* je vhodná na segmentáciu rukopisných dokumentov, *Printed Block Detection* na segmentáciu tlačených dokumentov.
2. označiť strany, na ktorých chcete automatickú segmentáciu vykonať:
   a. na jednej strane (*Current page*) – táto voľba je automaticky prednastavená,
   b. na celom, resp. len určitých stranách dokumentu (*Pages*) – po kliknutí na krúžok pred označením strán sa otvorí okienko na zápis rozsahu strán, prípadne rozsah strán vyberte kliknutím na tri bodky za okienkom na zápis rozsahu.

   Na začiatok odporúčame spustiť segmentáciu na jednej strane, aby ste si overili, či je automatická segmentácia pre váš typ dokumentu vyhovujúca.
3. vybrať objekty segmentácie:
   a. na segmentáciu textových rámcov zakliknite štvorček *Find Text-Regions*,
   b. na segmentáciu riadkov zakliknite štvorček *Find Lines.*

   Segmentáciu oboch objektov môžete urobiť súčasne.

Segmentáciu spustíte kliknutím na tlačidlo Spustiť (*Run*). Zobrazí sa dialógové okno s nastavenými parametrami segmentácie, ktoré potvrďte tlačidlom OK.

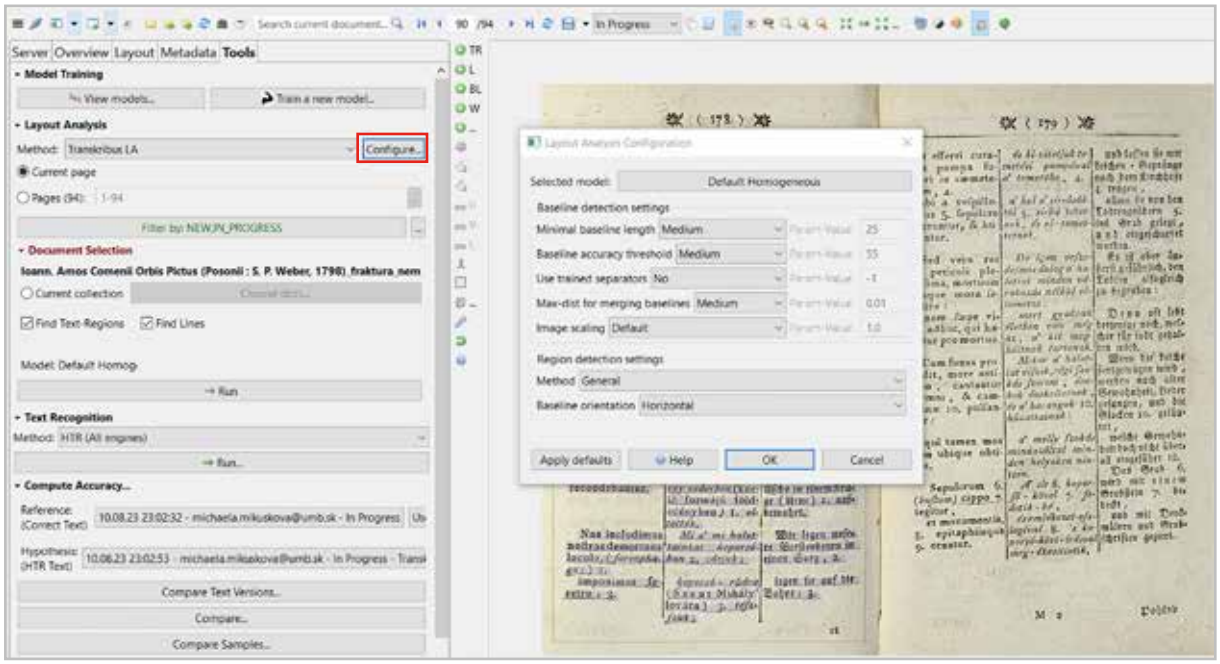


Obrázok 58 Dôležité prvky nastavenia automatickej segmentácie (Method Transkribus LA)

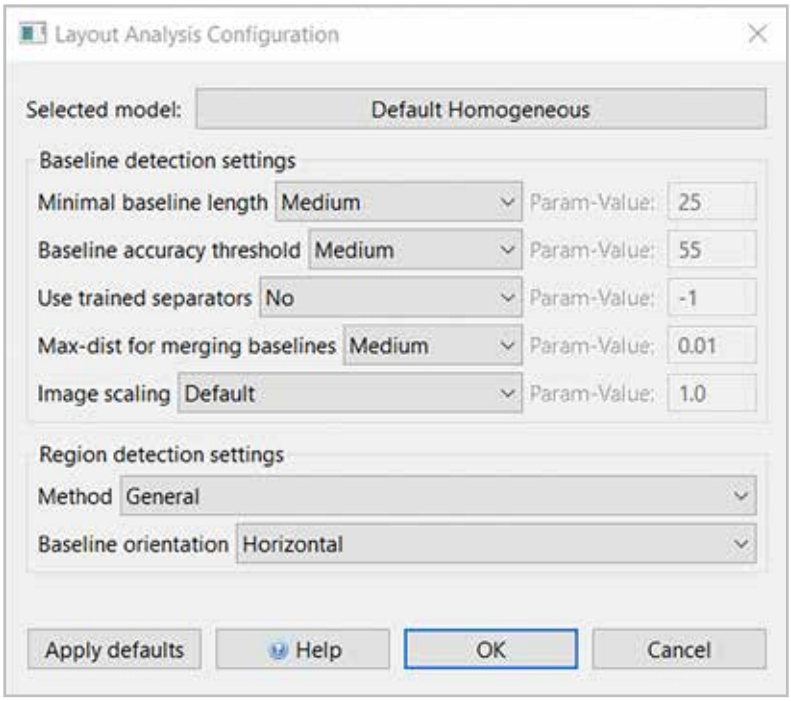
4.1.1.1. Pokročilé nástroje na nastavenie automatickej segmentácie textu

Proces automatickej segmentácie je defaultne nastavený a nemusí vyhovovať každému dokumentu. Používatelia Transkribus expert klienta majú k dispozícii nástroje na úpravu predvolených parametrov. Dialógové okno s ponukou sa otvorí po kliknutí na ikonu Nastaviť (*Configure…*) v sekcii Analýza rozloženia (*Layout analysis*). Nastavenie pozostáva z dvoch krokov:

1. výber modelu,
2. úprava parametrov objektov segmentácie.



Obrázok 59 Otvorenie nástrojov konfigurácie automatickej segmentácie

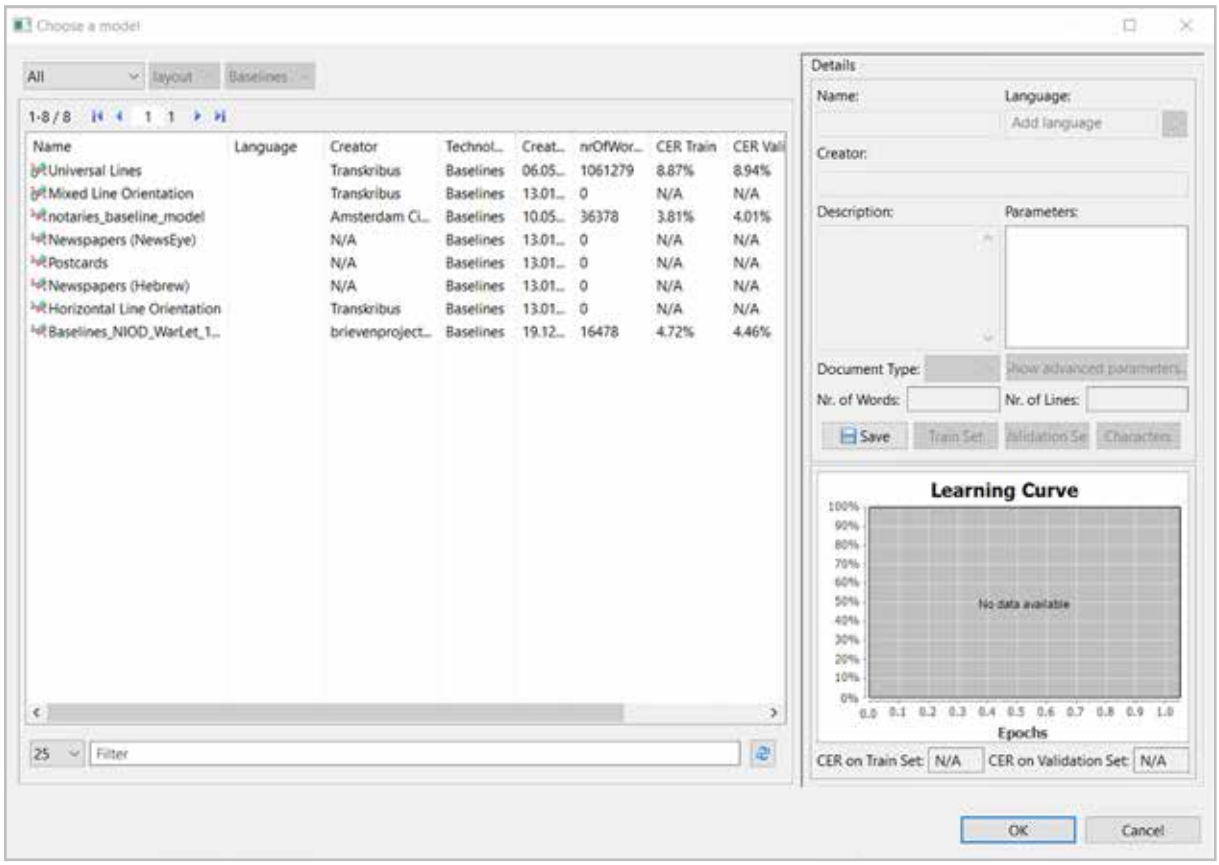


Obrázok 60 Detail dialógového okna pre nastavenie parametrov pokročilej segmentácie

Výber modelu
Na výber je niekoľko vytrénovaných modelov rozloženia obsahu:

- v dialógovom okne kliknite na *Default Homogenous*, otvorí sa ponuka v výberom modelov *Choose model*,
- k dispozícii je niekoľko modelov, ktoré vytrénovali vývojový tím softvéru Transkribus alebo používateľská komunita:
  - *Universal Lines* – najvšeobecnejší model, ktorý je v súčasnosti na platforme k dispozícii. Tento model odporúčame použiť, pokiaľ si nie ste istí výberom optimálneho modelu, ktorý bude vyhovovať vlastnostiam dokumentu, s ktorým pracujete,
  - *Mixed Line Orientation* – model pre rôznorodé rozloženie textu na snímkach, t. j. text je písaný vo viacerých smeroch,
  - *Horizontal Line Orientation* – model pre dokumenty s homogénnym rozložením textu, t. j. len horizontálne alebo vertikálne čiary.

K dispozícii sú aj modely zohľadňujúce štruktúru novín, pohľadníc a modely vytrénované pre špecifickú typológiu dokumentu.



Obrázok 61 Ponuka modelov na segmentáciu dokumentu

Úprava parametrov objektov segmentácie
Nastavenie parametrov segmentácie možno vykonávať pre oblasť základných čiar a textových rámcov.

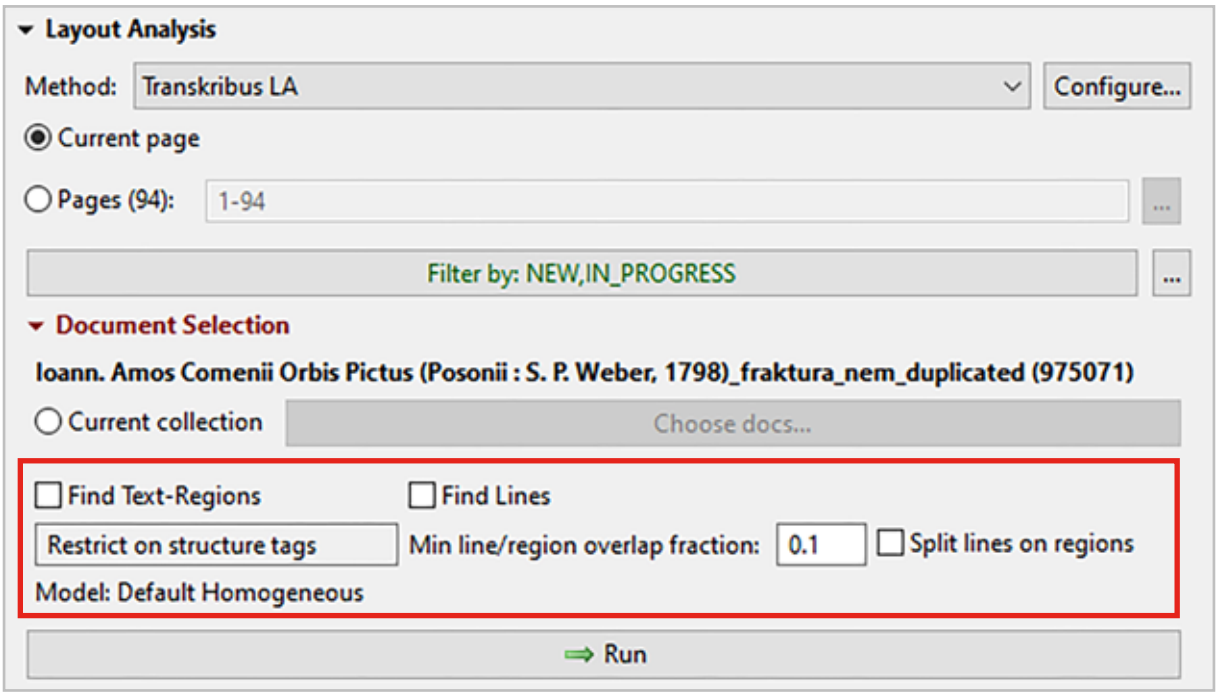
Parametre analýzy rozloženia (*Layout analysis*) ***základných čiar*** (*Baselines*)
Úpravu prednastavených hodnôt odporúčame, ak pri segmentácii bolo rozpoznaných príliš málo/veľa základných čiar alebo ak boli nesprávne spojené/oddelené. Pre každý parameter môžete vybrať jednu z troch navrhovaných hodnôt – nízka (*Low*), stredná (*Medium*), vysoká (*High*) alebo si hodnotu prispôsobte (*Custom*):

- minimálna dĺžka základnej čiary (*Minimal baseline length*) – udáva sa v pixeloch. Ak algoritmus v procese segmentácie detekuje základné čiary pod nastavenou dĺžkou, vynechá ich,
- prahová hodnota presnosti základnej čiary (*Baseline accuracy threshold*) – v prvej fáze rozpoznávania rozloženia sa každý pixel označí ako základná čiara, oddeľovač alebo iné. Prah presnosti základnej čiary sa pohybuje v rozmedzí od 0 do 255. Vyššie hodnoty sa prejavia vo väčšej presnosti rozpoznaných základných čiar. Pri obrázkoch s nižším rozlíšením sa pri neúspešnej detekcii základných čiar odporúča hodnoty znížiť.
- použitie natrénovaných oddeľovačov (*Use trained separators*) – oddeľovače sú malé zvislé čiary nakreslené vedľa každej základnej čiary, označujú jej začiatok a koniec. Rozpoznávajú sa v prvej fáze analýzy rozloženia. Prahová hodnota oddeľovača sa pohybuje v rozmedzí od 0 do 255. 0 znamená, že oddeľovače sa vôbec nepoužívajú. Zvyčajne aj nižšie hodnoty zabránia spájaniu základných čiar. Použite napr. hodnotu 1, ak chcete informácie o oddeľovačoch používať niekedy (*Sometimes*) a vyššie hodnoty, ak ich chcete používať stále (*Always*).
- maximálna vzdialenosť na zlučovanie (*Max-dist for merging*) – v druhej fáze sa softvér pokúša zlúčiť blízke základné čiary za predpokladu, že je ich vzdialenosť menšia ako nastavená hodnota. Použite hodnotu *Low* na zlúčenie čiar, ktoré sa na dokumente nachádzajú bližšie ako 0,5 % šírky obrazu, *Medium* na zlúčenie čiar, ktoré sú bližšie ako 1 % šírky obrazu, alebo *High* na zlúčenie čiar, ktoré sú od seba vzdialené viac ako 1 %, ale bližšie ako 5 % šírky obrazu. Vo väčšine prípadov by mala dobre fungovať voľba *Medium*.
- škálovanie obrázka (*Image scaling*) – môžete sa rozhodnúť, či chcete zvýšiť škálovanie obrázkov s nízkym rozlíšením alebo znížiť škálovanie obrázkov s vysokým rozlíšením. Túto funkciu odporúčame vyskúšať len vtedy, keď segmentácia s predvolenými nastaveniami nefunguje, napr. detekuje žiadne/málo základných čiar.

Parametre nastavenia generovania textových rámcov (*Text regions*)
Po analýze riadkov a základných čiar dochádza ich k zoskupeniu do blokov. K dispozícii sú dve metódy zhlukovania:

- všeobecná (*General*) – zhlukuje riadky zľava doprava. S nastavením tejto hodnoty súvisí aj nastavenie orientácie základných čiar (*Baseline orientation*). Nastavte hodnotu *Horizontal*, ak sa v dokumente nachádzajú len horizontálne orientované riadky, alebo hodnotu *Mixed*, ak sú v dokumente aj riadky otočené o 0, 90, 180 a 270 stupňov.
- vlastná (*Custom*) – ide o jednoduché aglomeratívne zhlukovanie založené na najľavejšom bode každého riadku. Zhlukuje čiary na základe ich vzdialenosti. Môžete nastaviť, či na snímke má byť jeden textový rámec (*One*), niekoľko (*Few*), stredne veľa (*Medium*), veľa (*Many*), alebo ich počet voliteľne prispôsobte (*Custom*).

Nastavenie parametrov ukončite v záložke **Tools**. Nezabudnite, že ak ste si vopred označili textové rámce, nesmie byť zaškrtnutý štvorček *Find Text-Regions* (viac v kapitole *1.1.3. Manuálna segmentácia*).



Obrázok 62 Ukončenie nastavenia parametrov segmentácie

4.1.1.2 Automatická segmentácia a rozpoznávanie textu

Automatickú analýzu rozloženia (*Layout analysis*) a transkripciu dokumentu môžete vykonať v jednom kroku. Slúži na to sekcia **Rozpoznávanie textu** (*Text recognition*), ktorú nájdete v záložke **Nástroje** (*Tools*). Pri transkripcii dokumentu týmto spôsobom treba aplikovať niektorý z vytrénovaných textových modelov.

Textový model je algoritmus umelej inteligencie vycvičený na určitom počte údajov (obrázkov a prepisov), ktorý dokáže zistiť najpravdepodobnejšiu postupnosť znakov pre každý segmentovaný riadok textu. Všeobecný model pre všetky rukopisy neexistuje, preto musíte vybrať čo najvhodnejší model pre písmo a jazyk dokumentu, s ktorým pracujete.

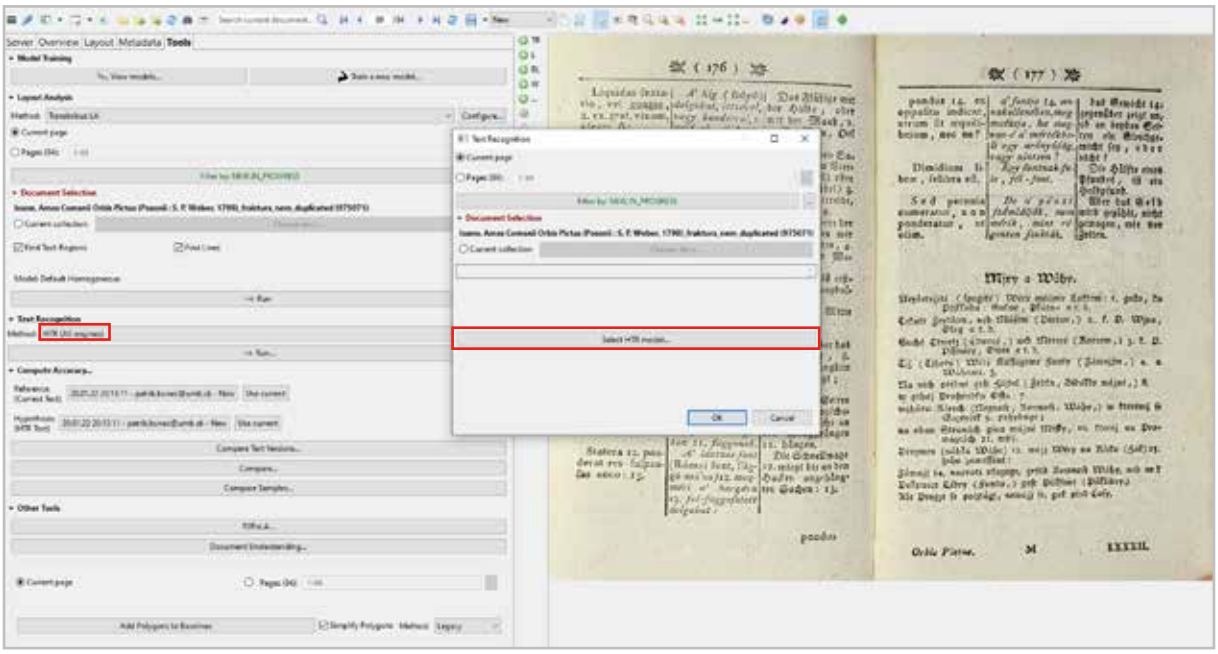
V rámci programu Transkribus je k dispozícii niekoľko verejných modelov, ktoré sprístupnila komunita a tím vývojárov programu Transkribus, aj súkromné modely, ktoré vytrénovali samotní používatelia platformy. Pred spustením rozpoznávania textu a segmentácie treba:

1. nastaviť metódu (*Method*) – automaticky býva prednastavená HTR (*All engines*), t. j. prehľadávanie všetkých dostupných modelov. K dispozícii je aj Transcribus OCR. Dvojitým kliknutím na *HTR (All engines)* sa otvorí dialógové okno.
2. označiť strany, na ktorých chcete automatickú segmentáciu vykonať:
   a. na jednej strane (*Current page*) – táto voľba je automaticky prednastavená,
   b. na celom, resp. len určitých stranách dokumentu (*Pages*) – po kliknutí na krúžok pred označením strán sa aktivuje okienko na zápis rozsahu strán, prípadne rozsah strán vyberte kliknutím na tri bodky za okienkom na zápis rozsahu,
3. vybrať model – modely môžete prehľadávať a filtrovať podľa metódy, ktorá bola použitá na vytrénovanie modelu, podľa jazyka, názvu, typu dokumentu, úspešnosti a i.

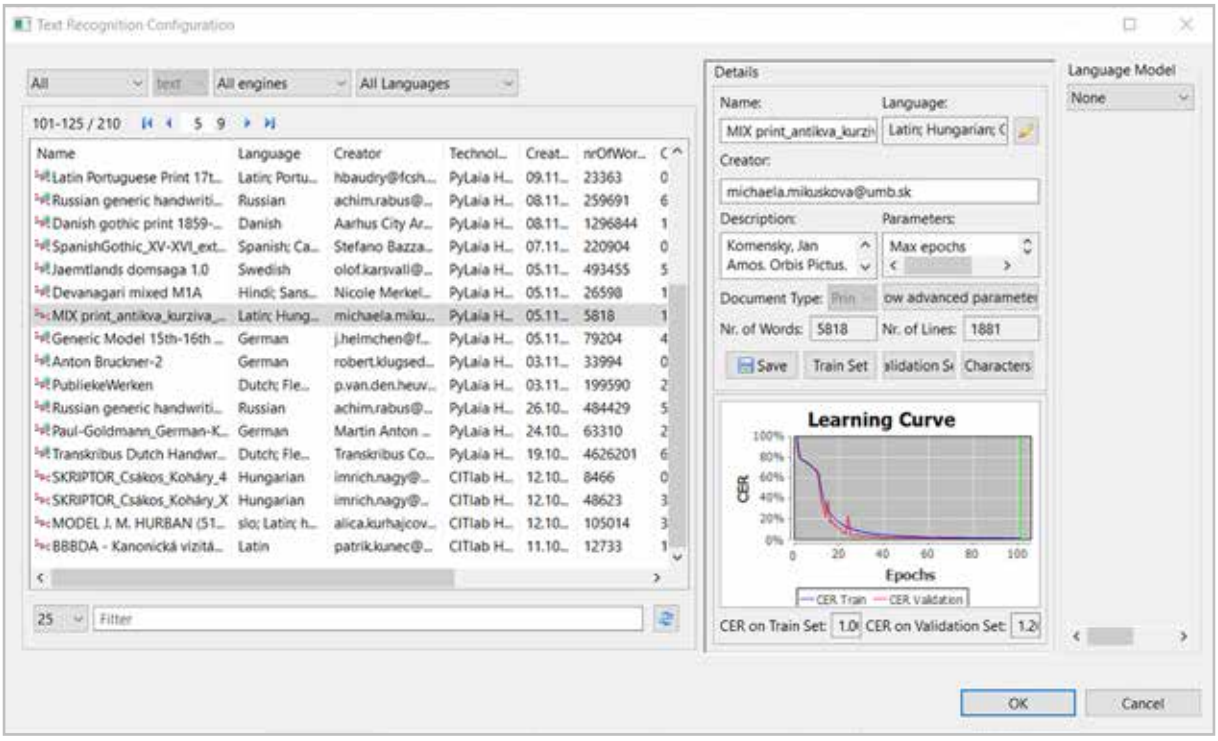
Nastaviť môžete aj definíciu riadkov a polygónov zaklinutím *Compute line polygons*.

Nastavenia potvrďte kliknutím na OK.

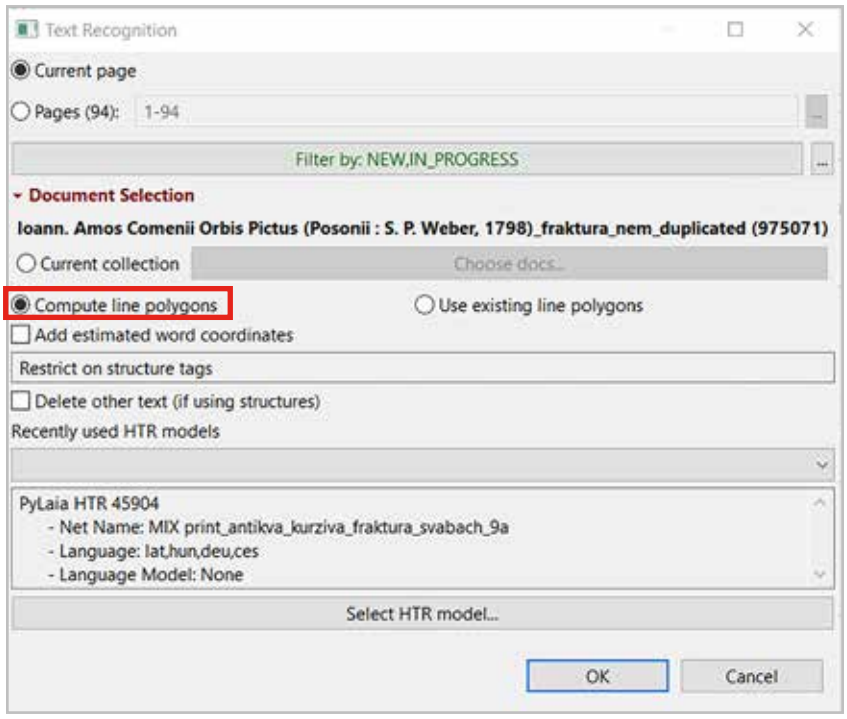
Rozpoznávanie textu a segmentáciu spustíte kliknutím na tlačidlo Spustiť (*Run*).



Obrázok 63 Výber modelu



Obrázok 64 Dialógové okno na výber modelu na rozpoznávanie textu



Obrázok 65 Dialógové okno na ďalšie nastavenia rozpoznávania textu a automatickej segmentácie

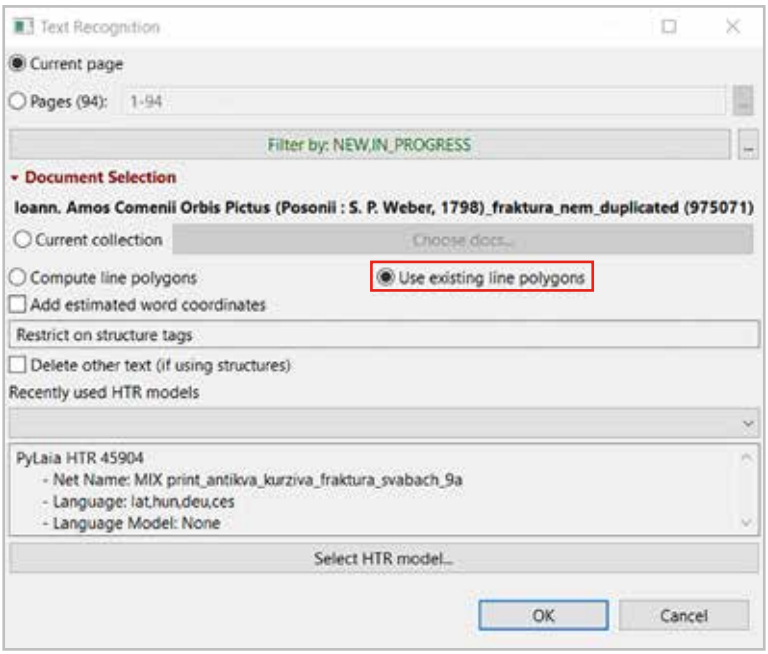
Po spustení rozpoznávania textu môžete skontrolovať stav riešenia zadanej úlohy cez ikonku (*Jobs*) v hlavnom menu Transkribus expert klient. Keď program segmentáciu ukončí, klient vás vyzve na opätovné načítanie stránky. Na snímke dokumentu sa zobrazia výsledky segmentácie a v textovom editore pod snímkou výsledky transkripcie.

4.1.1.3 Generovanie textu v dokumentoch so zložitou štruktúrou

V prípade, že pracujete s dokumentom so zložitejšou štruktúrou, napr. text obsahuje tabuľky, marginálie, viacero stĺpcov, odporúčame:

1. vykonať analýzu rozmiestnenia (*Layout Analysis*) objektov segmentácie – textové rámce, riadky,
2. skontrolovať poradie čítania objektov segmentácie (viac v kapitole *4.2 Kontroly po automatickej a manuálnej segmentácii*),
3. spustiť funkciu rozpoznávania textu (*Text Recognition*).

Analýzu rozloženia, t. j. segmentáciu objektov ste vykonali v prvom kroku, preto je dôležité, aby voľba *Use existing line polygons* bola zakliknutá. Zabránite tým opätovnej analýze rozloženia obsahu.



Obrázok 66 Dialógové okno na ďalšie nastavenia rozpoznávania textu v manuálne nasegmentovanom dokumente

4.1.2 Manuálna segmentácia

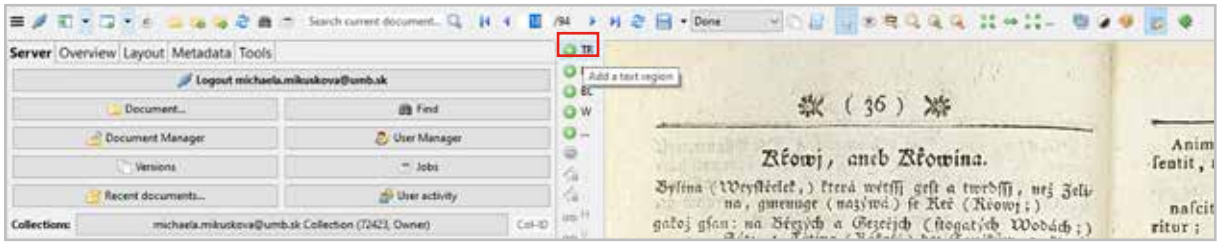
Pri manuálnej segmentácii je proces analýzy snímok rozdelený do dvoch krokov:

1. manuálne označenie textových rámcov,
2. spustenie automatickej segmentácie riadkových rámcov a základných čiar v manuálne označených textových rámcoch.

Treba starostlivo zvážiť členenie textu a jeho rozdelenie na textové rámce. Počet rámcov na jednej strane dokumentu závisí od jeho štruktúry a obsahu.

Manuálne označenie textových rámcov

Nástroje na vytváranie textových rámcov a iné úpravy segmentácie nájdete v editore *Canvas* naľavo od snímky dokumentu. Na vytvorenie textového rámca kliknite na ikonku TR (*Add a text region*).



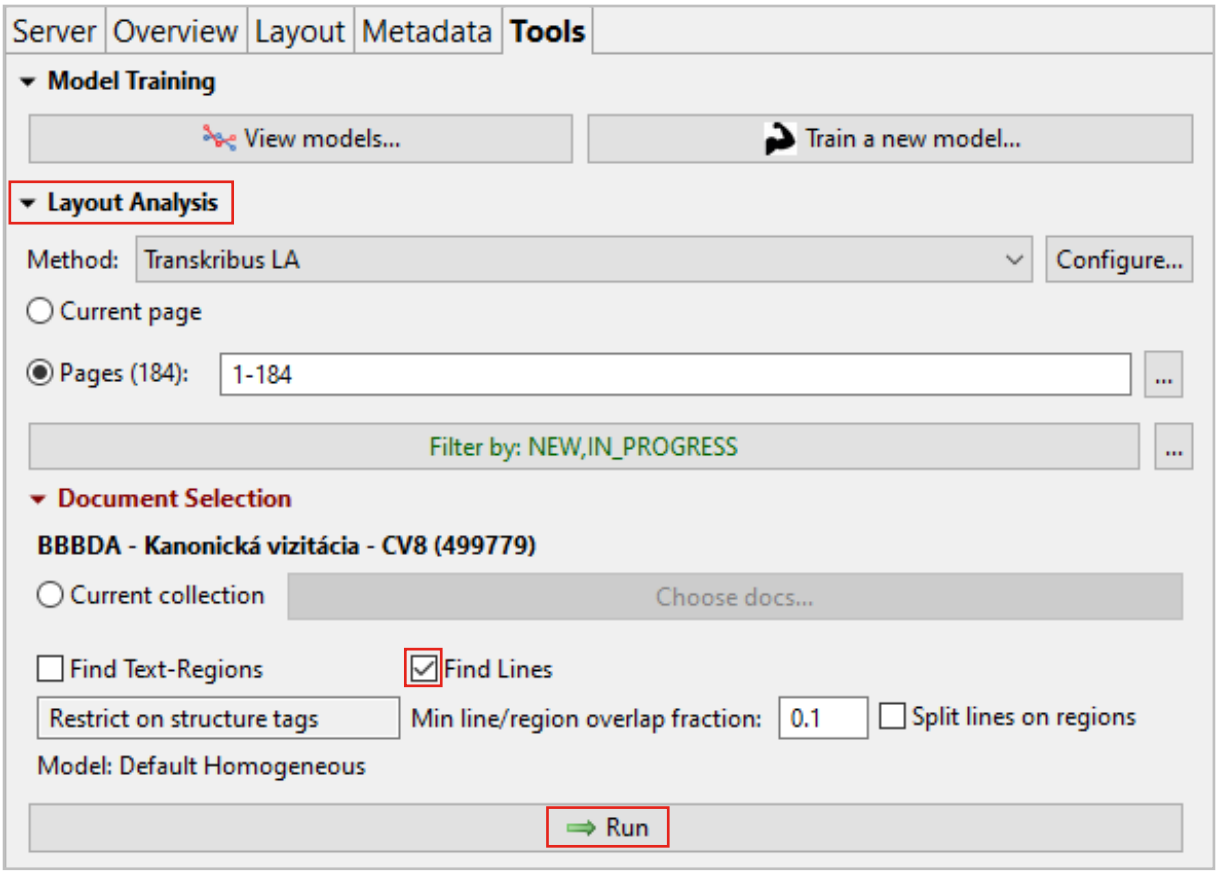
Obrázok 67 Výber funkcie Add a text region

Textové rámce majú zvyčajne tvar štvoruholníka (štvorca alebo obdĺžnika). Na snímke kurzorom kliknite na miesto, z ktorého chcete textový rámec začať vytvárať, t. j. na oblasť, kde sa bude nachádzať jeden z vrcholov štvoruholníka. Z tohto miesta druhým kliknutím a postupným ťahaním po snímke označte priestor, ktorý bude vymedzovať príslušný textový rámec. Hranice vytvoreného textového rámca označujú zelené čiary. Tie môžete ďalej upravovať a posúvať.

Nesprávne vytvorený textový rámec môžete vymazať (viac sa dočítate v kapitole *4.2 Opravy po automatickej a manuálnej segmentácii*). Pri prechode na inú stránku dokumentu zmeny uložte.

Spustenie automatickej segmentácie riadkov a základných čiar

Keď máte vymedzené textové rámce, môžete pristúpiť k automatickej segmentácii riadkov. Otvorte záložku **Nástroje** (*Tools*) na ľavej strane klienta pod hlavným menu. Prejdite do sekcie **Analýza rozloženia** (*Layout Analysis*).



Obrázok 68 Dôležité prvky nastavenia manuálnej segmentácie

Pred spustením segmentácie treba:

1. nastaviť metódu (*Method*) – automaticky býva prednastavená *Transkribus LA*. Kliknutím na šípku na konci riadka sa otvoria ďalšie metódy (*Kraken*, *Printed Block Detection* a *Separator Detection*). Voľbu metódy vyberte podľa typu dokumentu, s ktorým pracujete. *Transkribus LA* je vhodná na segmentáciu rukopisných dokumentov, *Printed Block Detection* na segmentáciu tlačených dokumentov.
2. označiť strany, na ktorých chcete automatickú segmentáciu vykonať:
   a. na jednej strane (*Current page*) – táto voľba je automaticky prednastavená,
   b. na celom, resp. len určitých stranách dokumentu (*Pages*) – po kliknutí na krúžok pred označením strán sa aktivuje okienko na zápis rozsahu strán, prípadne rozsah strán vyberte kliknutím na tri bodky za okienkom na zápis rozsahu.

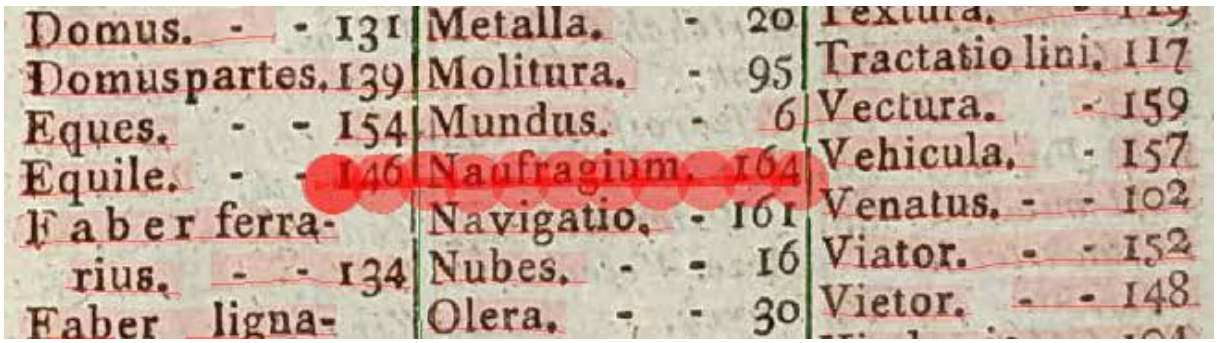
3. vybrať objekty segmentácie:
    a. textové rámce už máte definované, preto je dôležité, aby štvorček *Find Text-Regions* zostal neoznačený,
    b. na segmentáciu riadkov kliknite na štvorček s popisom *Find Lines.*

Segmentáciu je možné obmedziť len na určitý typ štruktúry obsahu kliknutím na *Restrict on structure tags* (viac o štrukturálnych metadátach v kapitole *7.2 Štrukturálne tagy*).

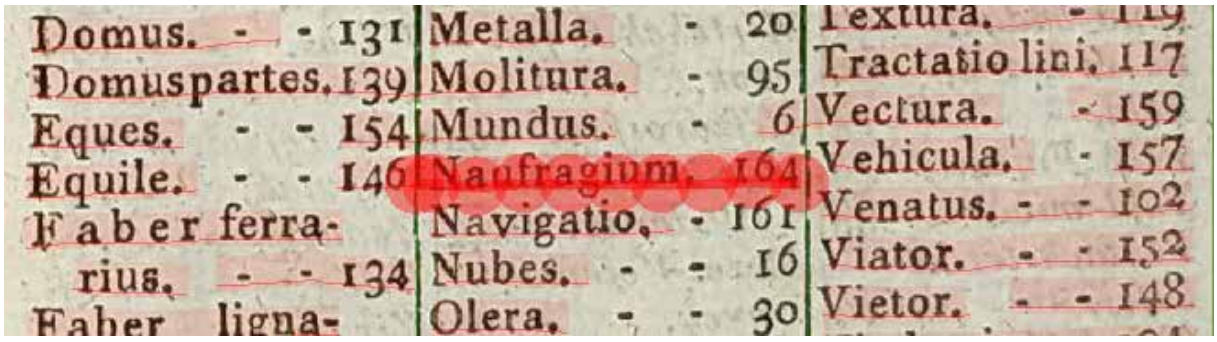
Nastaviť môžete podiel prekrytia základných čiar riadkov a textových rámcov nastavením hodnoty *Min. line/Region overlap fraction*. Ak chcete, aby základné čiary prekrývali hranice textových rámcov, treba hodnotu zvýšiť.

Zakliknutie funkcie *Split lines on regions* zabezpečí, že sa riadky striktne riadia hranicou textových rámcov. Toto nastavenie je veľmi užitočné pri segmentácií dokumentov, v ktorých sa textové rámce nachádzajú tesne vedľa seba a text v nich sa končí takmer na hranici textového rámca. Zabráni sa spojeniu riadkov do jedného dlhého riadka (príklady na obrázkoch nižšie).

Segmentáciu spustíte kliknutím na tlačidlo Spustiť (*Run*). Zobrazí sa dialógové okno s nastavenými parametrami segmentácie, ktoré potvrďte tlačidlom OK.



Obrázok 69 Výsledok automatickej segmentácie bez použitia funkcie Split lines on regions – riadky v dvoch susediacich textových rámcoch sú spojené



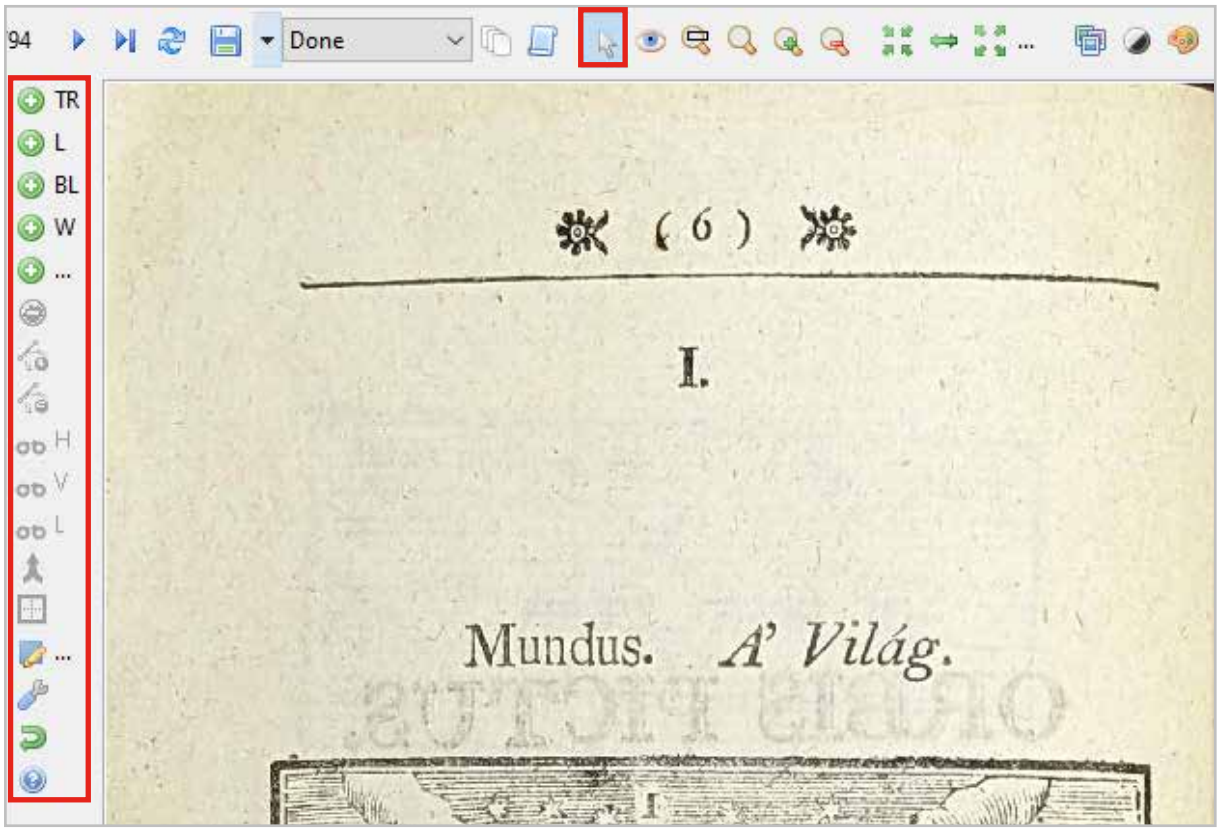
Obrázok 70 Výsledok automatickej segmentácie s použitím funkcie Split lines on regions – riadky v dvoch susediacich textových rámoch sú oddelené

Po spustení segmentácie môžete skontrolovať stav riešenia zadanej úlohy cez ikonku *Jobs* v hlavnom menu Transkribus expert klienta. Keď program segmentáciu ukončí, klient vás vyzve na opätovné načítanie stránky. Na stránke sa zobrazia výsledky segmentácie.

4.2 Opravy po automatickej a manuálnej segmentácii

Výsledky manuálnej a automatickej segmentácie nie sú vždy ideálne a vo väčšine prípadov je potrebné urobiť ďalšie opravy. Časová náročnosť úprav závisí od štruktúry dokumentu a správne zvoleného typu analýzy rozloženia textu, t. j. segmentácie.

Nástroje na korekciu segmentácie sú dostupné v editore *Canvas* na ľavej strane snímky dokumentu, s ktorým pracujete. Funkcie sú reprezentované ikonkami, ktorých popis sa zobrazí, keď na príslušnú ikonku prejdete kurzorom. Zvolený nástroj v *Canvas* sa deaktivuje po vykonaní požadovaného úkonu, alebo kliknutím na ikonku (*Selection mode*) v hlavnom menu, alebo stlačením klávesu ESC.



Obrázok 71 Ikony nástrojov editora Canvas

4.2.1 Korekcia textových rámcov (*Text Regions*)

Textové rámce, ktoré sa pri automatickej a manuálnej segmentácií vytvárajú, majú tvar štvorca, alebo obdĺžnika v závislosti od textu, ktorý označujú. Mali by obklopovať celý text, ktorý je obsiahnutý na snímke dokumentu a má byť predmetom transkripcie. Počet a štruktúra textových rámcov závisí od štruktúry a obsahu dokumentu. Pri manuálnej segmentácii je niekedy potrebné vytvoriť špecifické typy a tvary textových rámcov. Aj pri automatickej segmentácii textových rámcov môžu nastať prípady, keď je nutné urobiť čiastočné korekcie. Editor *Canvas* poskytuje niekoľko nástrojov na prácu s textovými rámcami.

Prispôsobenie textového rámca
Štandardne sú hranice textového rámca na seba kolmé a definované štyrmi kontrolnými bodmi, ktoré vymedzujú vrcholy rámca. Textové rámce je možné prispôsobovať posúvaním kontrolných bodov, prípadne posúvaním čiar označujúcich hranice rámca.

Pri manuálnom vytváraní textových rámcov môžu nastať prípady, že sa rámce prekrývajú, alebo text z jedného rámca čiastočne prechádza do iného. Rámce je možné upravovať pomocou pridávania kontrolných bodov, čím sa vytvára polygón.

Na úpravu hraníc textových rámcov:

- v editore *Canvas* kliknite na ikonku (*Add point to selected shape*),
- na zelených čiarach označujúcich hranice textového rámca pridajte ďalšie kontrolné body,
- textový rámec pomocou pridaných bodov upravte na požadovaný tvar.

Rozdelenie textového rámca
Textové rámce je niekedy potrebné rozdeliť, lebo text, ktorý rámec označuje, navzájom nesúvisí, napr. hlavný text dokumentu od marginálnych poznámok.

Pre rozdelenie jedného textového rámca na dva, rámec označte kurzorom. Podľa toho, ako potrebujete rámec rozdeliť v editore *Canvas*, vyberte príslušnú ikonku:

- horizontálne rozdelenie – ikonku  (*Splits* a *shape with* a *horizontal line*),
- vertikálne rozdelenie – ikonka  (*Splits* a *shape with* a *vertical line*),
- prispôsobiteľné rozdelenie – ikonka  (*Splits* a *shape with* a *custom polyline*).

Po zvolení správnej funkcie v označenom textovom rámci kliknite kurzorom na miesto, kde ho chcete rozdeliť.

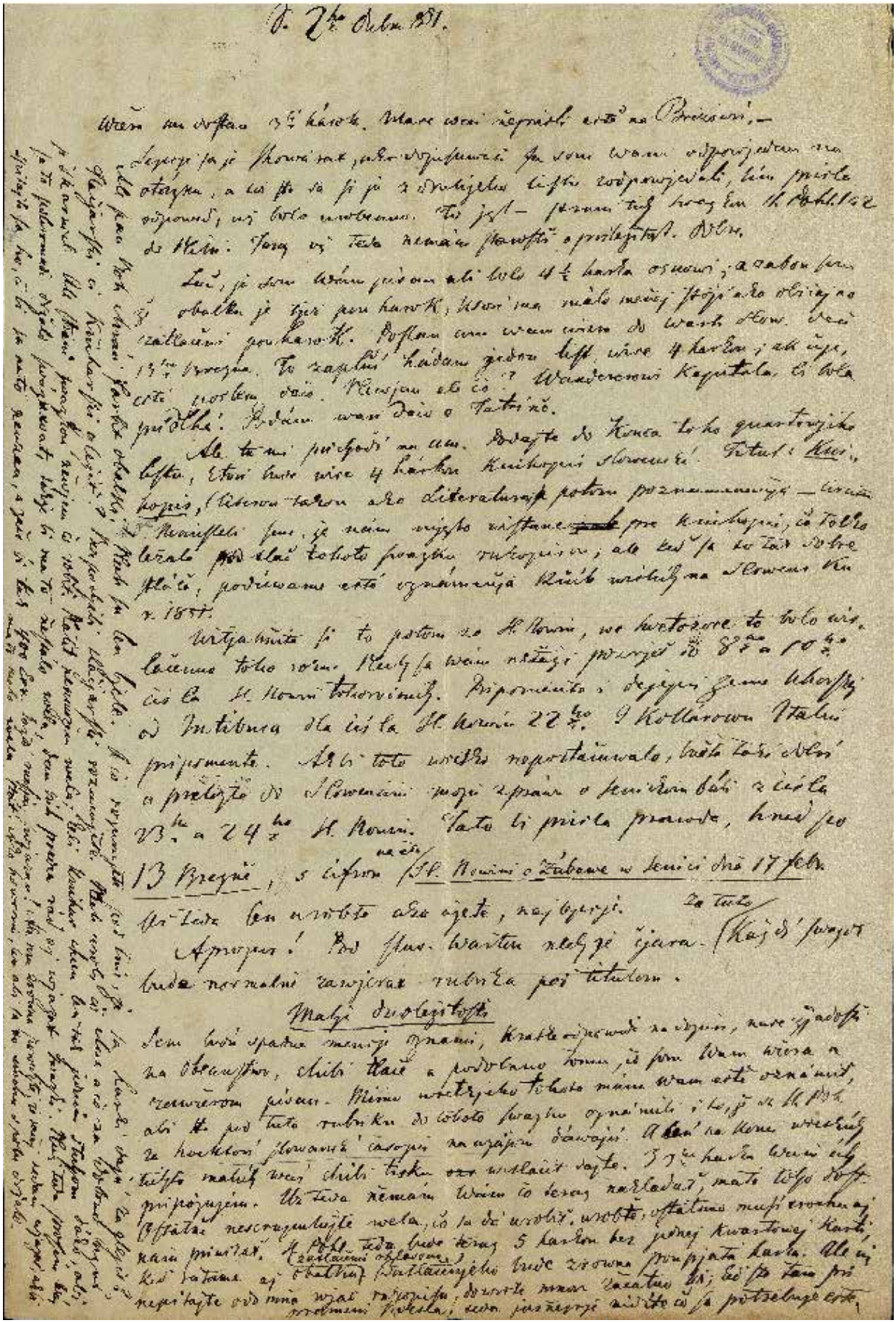
Spojenie textových rámcov
Automatickou segmentáciou môžu vzniknúť dva textové rámce, ktoré treba spojiť do jedného. Na spojenie viacerých rámcov:

- na klávesnici stlačte CTRL a kurzorom označte rámce, ktoré chcete spojiť,
- v editore *Canvas* kliknite na ikonku  (*Merges the selected shapes*).

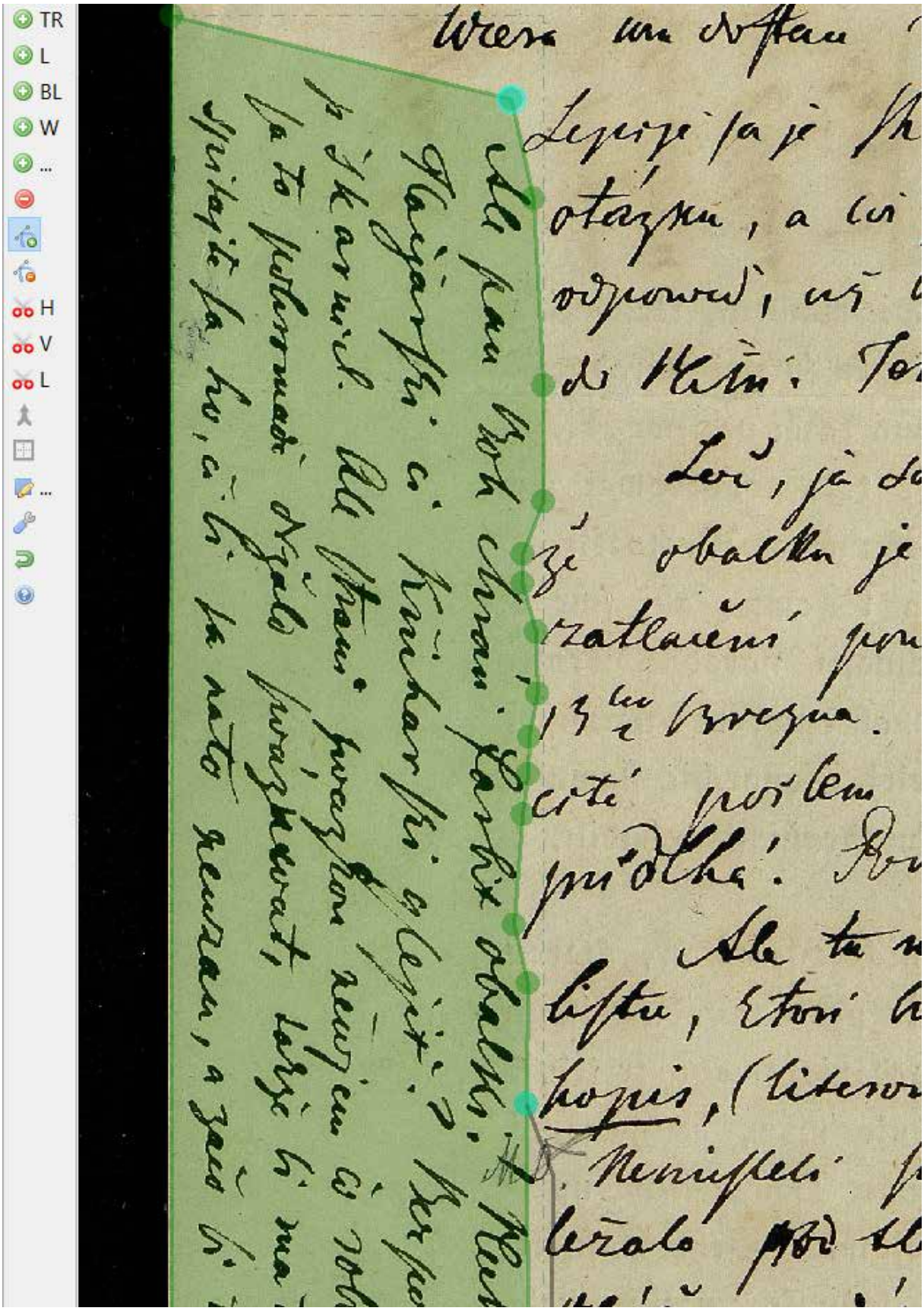
Odstránenie textových rámcov
Pri automatickej segmentácii môže vzniknúť nežiaduci textový rámec na mieste, kde sa nachádzajú rôzne šmuhy, text presvitá z inej strany a pod. Vyskytnúť sa môžu aj prípady, že v jednom textovom rámci vzniknú dva rámce. Tieto textové rámce treba odstrániť, aby nenarúšali štruktúru dokumentu, prípadne neoznačovali nežiaduce riadky, ktoré by mohli znižovať kvalitu vytrénovaného modelu. Na odstránenie rámca:

- kurzorom označte rámec, ktorý chcete vymazať,
- v editore *Canvas* kliknite na ikonku  (*Remove* a *shape*) alebo stlačte kláves DELETE.

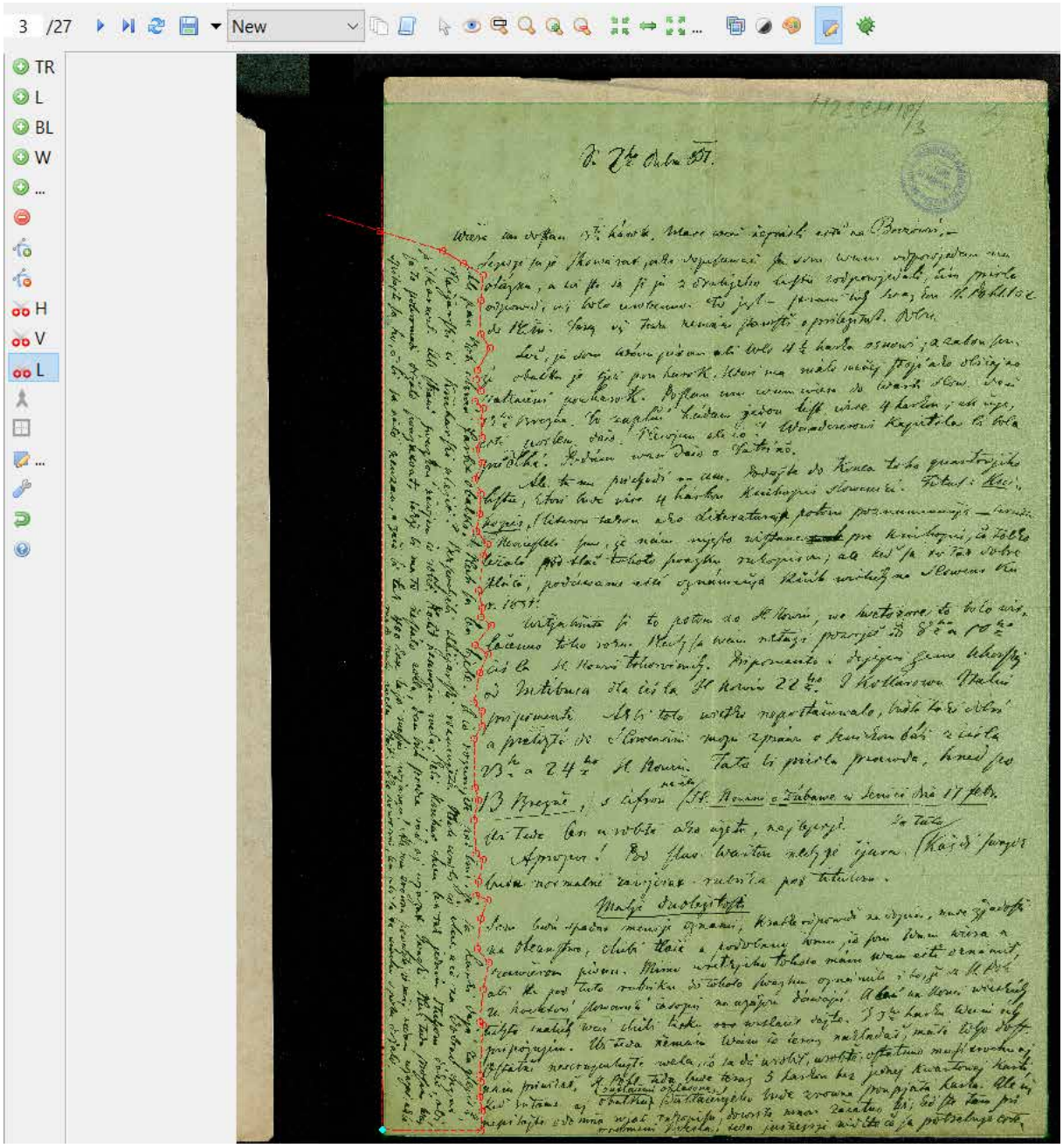
Ak odstraňujete rámec, v ktorom sú označené riadky a základné čiary, odstránia sa aj tie.



Obrázok 72 Kombinované horizontálno-vertikálne rozloženie textu vyžadujúce vytvorenie špecifických textových rámcov

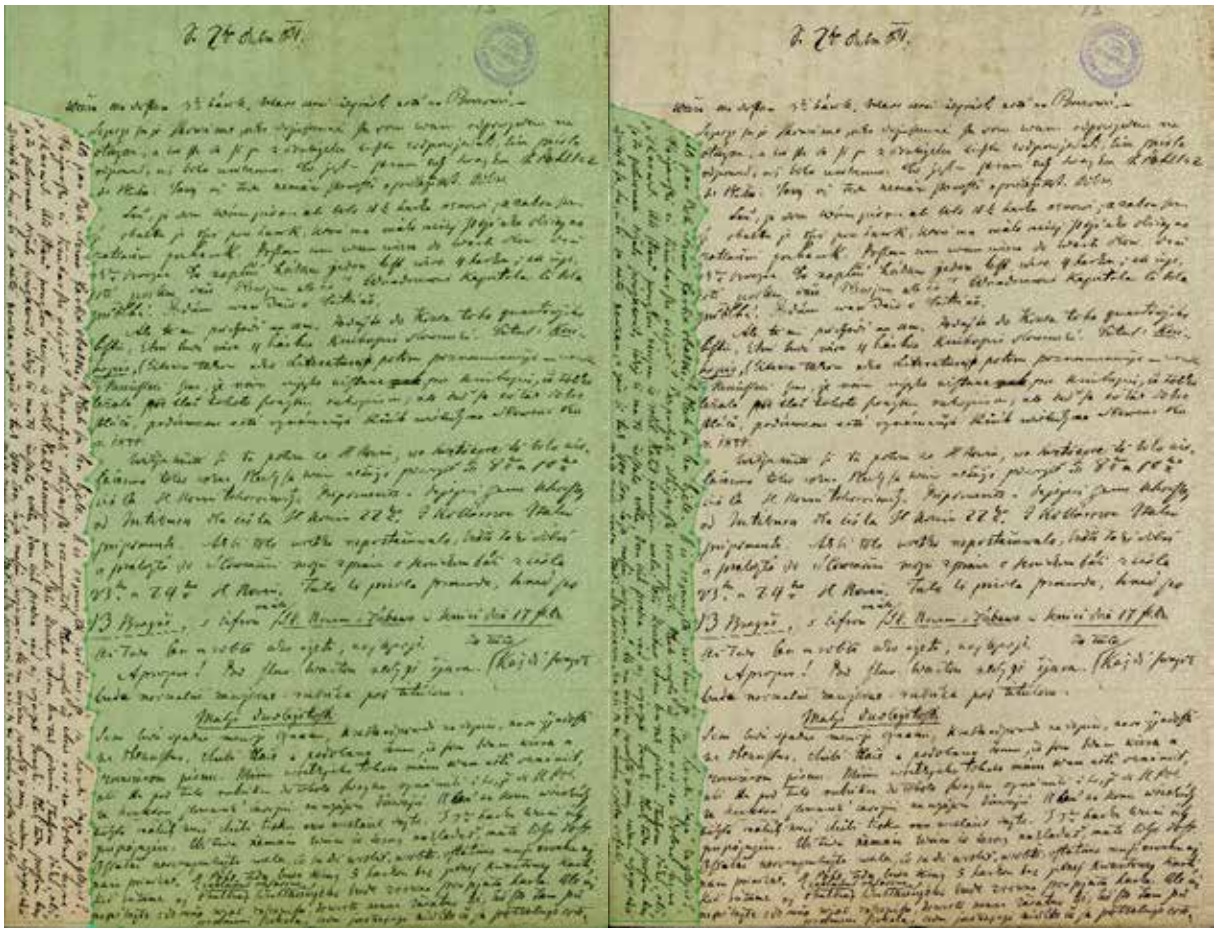


Obrázok 73 Detail manuálnej úpravy textových rámcov (polygónov) pridávaním nových bodov definujúcich hranice rámcov pomocou funkcie (Add point to selected shape). Vyžaduje tvorbu dvoch samostatných rámcov a vykresľovanie komplikovaného rozdelenia rámcov samostatne.



Obrázok 74 Manuálna úprava rámcov (polygónov) pomocou funkcie L
(Splits a shape with a custom polyline)

Na začiatku vytvoríte jeden veľký textový rámec zahŕňajúci celý text. Následne oddelíte horizontálny text od vertikálneho pridávaním samostatných bodov funkciou *Splits* a *shape with* a *custom polyline*, ktoré ukončíte spojením linky v mieste, kde ste začali. Súčasne sa vytvoria dva samostatné rámce.

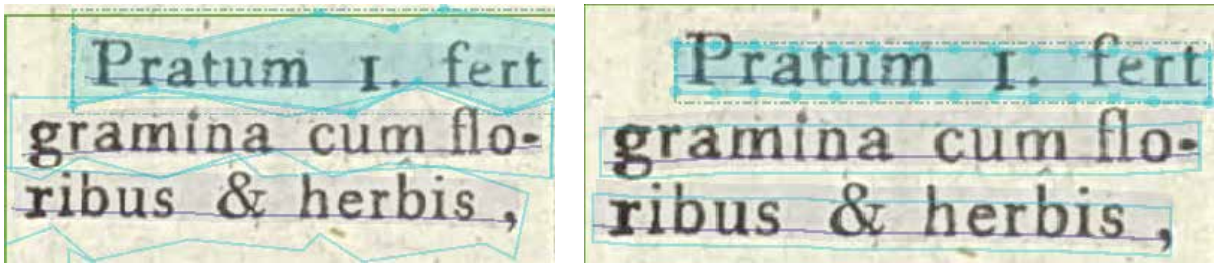


Obrázok 75 Výsledok rozdelenia rámcov s použitím funkcie  *(Splits a shape with a custom polyline) – dva samostatné rámce textu*

4.2.2 Korekcia riadkových rámcov (*Line Regions*)

Riadkové rámce sú viditeľné v profile *Default a Transcription*. Vymedzujú ich mnohouholníky, v ktorých sa nachádza ručne písaný alebo tlačený text príslušného riadku v textovom rámci. Na snímke dokumentu ich reprezentuje tenká čiara tyrkysovej farby. Táto čiara spája body, ktorých počet závisí od dĺžky textu nachádzajúceho sa v príslušnom riadku.

V procese transkripcie nemajú význam, preto ich netreba manuálne upravovať. Zmeny na úrovni riadka sa vykonávajú na úrovni základnej čiary (*Baseline*). Tieto zmeny sa následne prejavia v úprave riadkového rámca. K prispôsobeniu riadkového rámca dochádza po spustení trénovania modelu alebo transkripcii.



Obrázok 76 Zobrazenie riadkového rámca po segmentácii (vľavo) a po transkripcii (vpravo)

4.2.3 Korekcie základných čiar (*Baselines*)

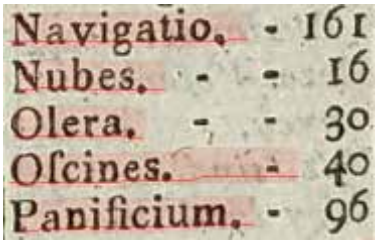
S chybami sa stretnete aj pri automatickej segmentácii riadkov a základných čiar. Základným referenčným bodom na rozpoznávanie textu je základná čiara (*Baseline*), ktorá popisuje polyčiaru tiahnucu sa pozdĺž spodnej časti riadku písaného alebo tlačeného textu. Jej úprave je preto potrebné venovať zvýšenú pozornosť.

Základná čiara je špecifikovaná červeným (v režime *Segmentation*) alebo fialovým (v režime *Default* a *Transcription*) označením.

Obrázok 77 Spôsob označenia základnej čiary v profile Segmentation (vľavo) a Transcription (vpravo)

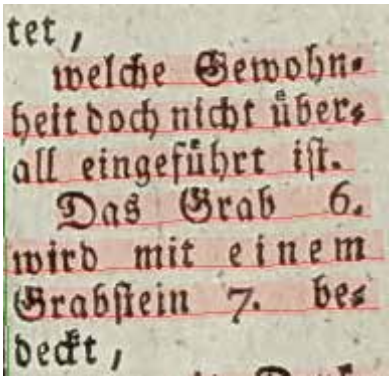
Najčastejšie sa môžete stretnúť s nasledujúcimi chybami pri analýze:

- základná čiara nekopíruje celý text v príslušnom riadku,



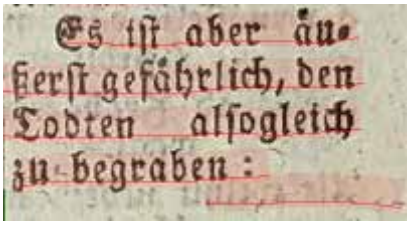
Obrázok 78 Nesprávne dotiahnutá základná čiara

- základná čiara sa nevytvorí tam, kde sa nachádza text dokumentu,



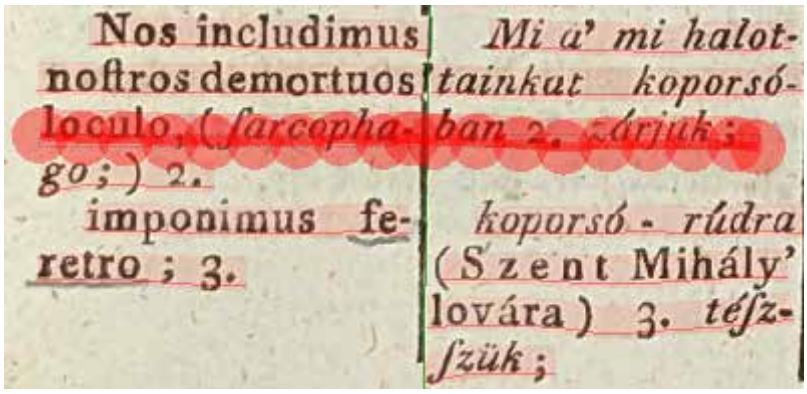
Obrázok 79 Nevytvorená základná čiara

- základná čiara sa vytvorí tam, kde sa text dokumentu nenachádza (napr. šmuha na dokumente, text presvitajúci z druhej strany papiera a pod.),



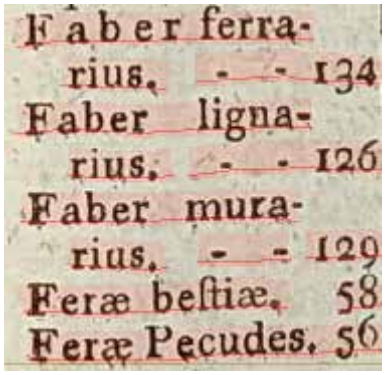
Obrázok 80 Základná čiara vytvorená na mieste, kde sa nevyskytuje text

- vytvorí sa jedna základná čiara cez viacero susediacich textových rámcov,



Obrázok 81 Základná čiara prechádzajúca cez dva rámce

- vytvorí sa viac základných čiar namiesto jednej.



Obrázok 82 Prerušovaná základná čiara na úrovni jedného riadku textu

Podobne ako na korekciu textových rámcov, aj na korekciu základných čiar používate nástroje v editore *Canvas*.

Úprava označenia základnej čiary

Keď na farebné označenie základnej čiary kliknete, zistíte, že ju tvorí niekoľko pospájaných kontrolných bodov. Začiatok a koniec základnej čiary nemusí presne zodpovedať textu. Prax ukazuje, že nie je nevyhnutné začiatok a koniec označenia doťahovať. Dôležité je, aby základná čiara správne kopírovala spodok riadku a písmená na nej „sedeli". Niekedy je však potrebné základnú čiaru upraviť, prípadne predĺžiť. Môžete tak urobiť dvomi spôsobmi:

1. natiahnutím okrajov základnej čiary v požadovanom smere:
   - kliknite na posledný bod základnej čiary,
   - posuňte ho do požadovanej strany.

2. pridaním nových bodov na základnej čiare:
   - v editore *Canvas* zvoľte na ikonku (*Add point to selected shape*),
   - kurzorom pridajte nový bod na požadované miesto a základnú čiaru upravte tak, aby kopírovala spodnú líniu písmen.

Pridanie základnej čiary
Ak sa pri segmentácii nevytvorila základná čiara tam, kde sa nachádza text:

- v editore *Canvas* vyberte ikonku  (*Add* a *baseline*),
- postupným klikaním kurzorom na spodnej línii písmen vložte niekoľko kontrolných bodov po celej dĺžke riadku,
- tvorbu základnej čiary ukončite dvojitým kliknutím, alebo stlačením klávesu ENTER v poslednom bode.

Základnú čiaru odporúčame označovať viacerými klikmi pozdĺž celého riadku tak, aby kopírovala písmená aj v prípade, že riadok nie je napísaný rovno.

Odstránenie základnej čiary
Na odstránenie prebytočného riadku:

- kurzorom označte základnú čiaru, ktorú chcete odstrániť,
- v editore *Canvas* kliknite na ikonku  (*Remove* a *shape*) alebo stlačte kláves DELETE.

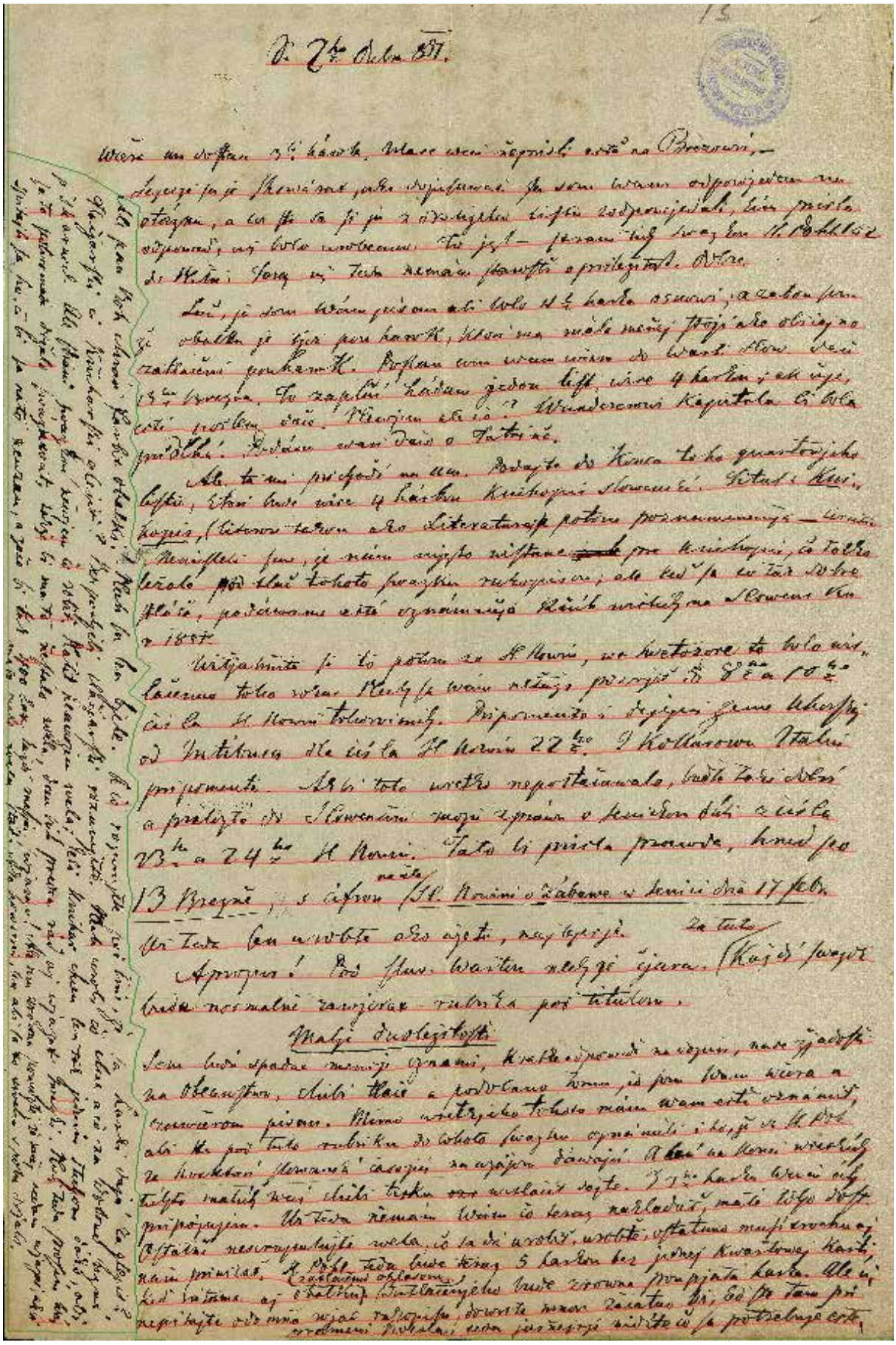
Rozdelenie základnej čiary
Ak potrebujete rozdeliť riadok, ktorý prechádza do viacerých textových rámcov:

- kurzorom označte základnú čiaru, ktorú chcete rozdeliť,
- v editore *Canvas* vyberte ikonku  (*Splits* a *shape with* a *horizontal line*),
- kurzorom kliknite na to miesto základnej čiary, kde je potrebné ju rozdeliť.

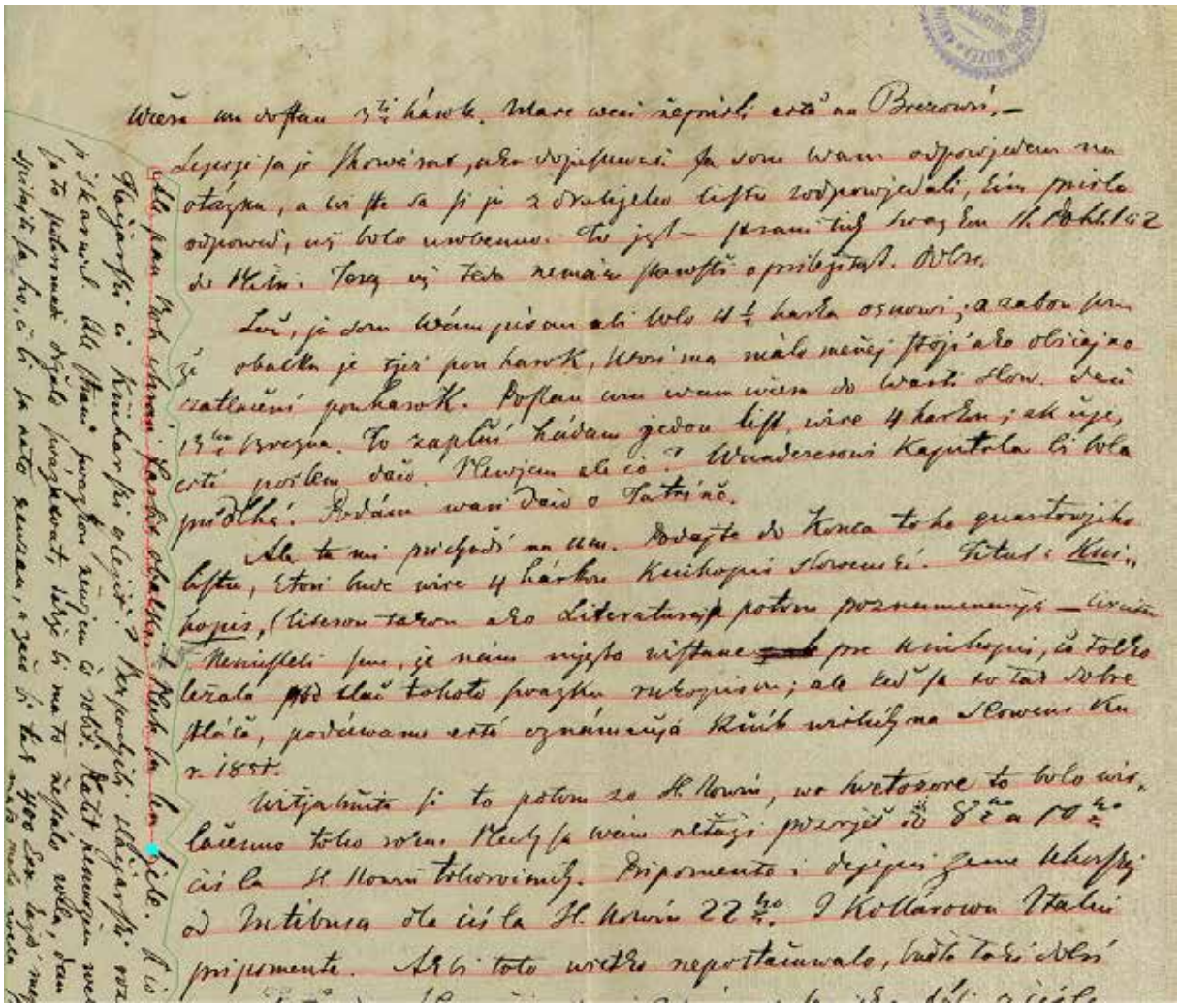
Spojenie základných čiar
Algoritmus niekedy nerozpozná štruktúru riadku a namiesto jedného riadku vytvorí dva, resp. aj viac. Na spojenie základných čiar:

- na klávesnici stlačte CTRL a kliknite na riadky, ktoré chcete spojiť,
- v editore *Canvas* vyberte ikonku  (*Merges the selected shapes*).

Základné čiary je možné zadefinovať aj vertikálne a kombinovať rôzne smery čiar na jednej strane dokumentu (napr. pri pohľadniciach alebo ako uvádza príklad nižšie).



Obrázok 83 Neuspokojivé výsledky automatickej segmentácie horizontálno-vertikálneho členenia textu



Obrázok 84 Manuálne doplnenie riadkov pri horizontálno-vertikálnom členení textu

4.2.4 Kontrola a úprava poradia čítania textových a riadkových rámcov

Mnohé dokumenty obsahujú nielen hlavný text, ale aj poznámky pod čiarou, marginálie, ktoré pridali iní používatelia dokumentu, prípadne je obsah dokumentu veľmi štruktúrovaný, napr. je zapísaný v stĺpcoch, obsahuje tabuľky a pod. Algoritmus pri analýze rozloženia usporadúva textové a riadkové rámce podľa ich grafického výskytu a automaticky ich čísluje podľa súradníc na snímke dokumentu, pričom postupuje od ľavého horného rohu smerom nadol.

Na trénovanie modelu nie je dôležité striktné poradie čítania textových rámcov a riadkov v nich. Toto poradie je však dôležité, ak chcete s textom následne pracovať, sprístupniť ho iným používateľom alebo ho publikovať. Na to, aby bol text s náročným rozložením usporiadaný pre čitateľa zrozumiteľne, má Transkribus expert klient k dispozícii nástroje, vďaka ktorým môžete zmeniť poradie čítania textových a riadkových rámcov a usporiadať ich do logického sledu.

Nástroje na úpravu poradia čítania textových a riadkových rámcov:

1. ikonka (*Shape visibility...*) v hlavnom menu – na korekcie menšieho rozsahu,
2. záložka Rozloženie (*Layout*).

Obrázok 85 Umiestnenie nástrojov na úpravu textových a riadkových rámcov

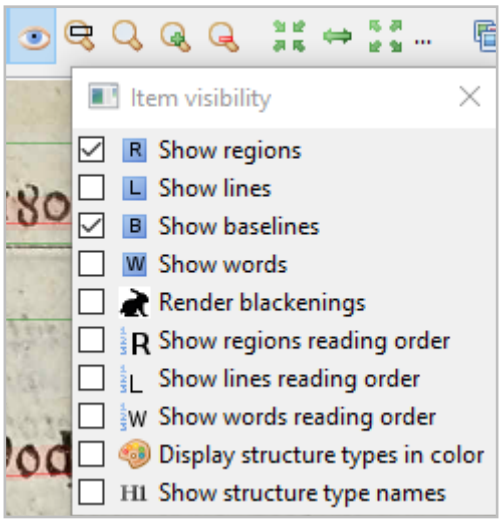
Pri dokumentoch so zložitým usporiadaním textu, kde sa poradie riadkov neriadi bežnými pravidlami, a pri dokumentoch, v ktorých ste vykonali veľa manuálnych opráv automatickej segmentácie, je možné oba nástroje kombinovať.

4.2.4.1 Viditeľnosť položky (*Item visibility*)

Táto funkcia slúži na opravu menších chýb poradia čítania objektov segmentácie.

Po kliknutí na ikonku (*Shape visibility*) sa otvorí okno Viditeľnosť položky (*Item visibility*), ktoré obsahuje možnosti pre zobrazenie jednotlivých objektov segmentácie:

- textových rámcov (*Show regions*),
- riadkov (*Show lines*),
- základných čiar (*Show baselines*),
- slov (*Show words*),
- začiernenia (*Render blackenings*),
- číselného označenia poradia čítania textových rámcov (*Show regions reading order*),
- číselného označenia poradia čítania riadkových rámcov (*Show lines reading order*),
- farebného rozlíšenia štrukturálnych tagov (*Display structure types in color*),
- pomenovania štrukturálnych tagov (*Show structure types names*).



Obrázok 86 Ponuka zobrazenia objektov segmentácie

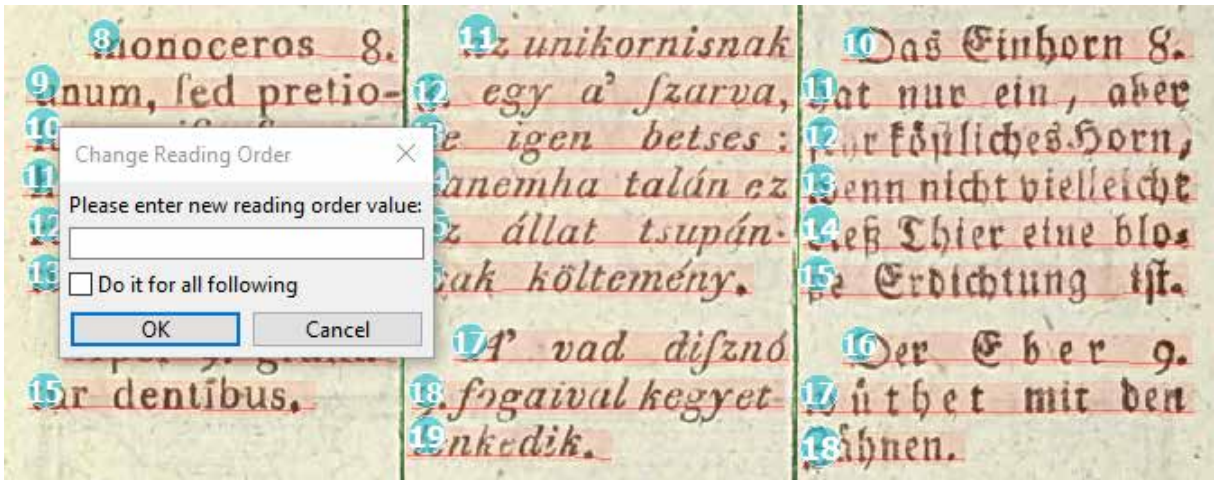
V režime *Segmentation* a *Transcription* býva automaticky nastavené zobrazenie textových polí a základných čiar. Ostatné možnosti si vyberáte podľa toho, aký objekt potrebujete zobraziť.

Každý textový rámec má vlastné číslovanie riadkov, t. j. prvý riadok textového rámca má byť označený číslicou jeden.

Obrázok 87 Zobrazenie správneho poradia čítania textových blokov a riadkov štruktúrovaného textu

Kontrola poradia čítania textových rámcov a riadkov:

- cez funkciu Viditeľnosť položky (*Item visiblity*) si zvoľte, ktorý objekt segmentácie (textový rámec alebo riadok) chcete zobraziť – najskôr odporúčame urobiť kontrolu poradia čítania textových rámcov a následne kontrolu poradia čítania riadkov v nich,
- na snímke dokumentu sa zobrazí číselné označenie poradia čítania príslušného objektu – čísla v zelenom krúžku označujú poradie textových rámcov, čísla v modrom krúžku označujú poradie riadkov v príslušnom textovom rámci,
- kliknite na číslo, ktoré potrebujete upraviť,
- zobrazí sa dialógové okno Zmeniť poradie čítania (*Change Reading Order*), do ktorého zapíšte novú, správnu hodnotu a potvrďte ju kliknutím na OK,
- zápisom novej hodnoty dôjde k prepisu nasledujúcich hodnôt daného objektu.



Obrázok 88 Dialógové okno na úpravu číselného poradia objektov

4.2.4.2 Záložka Rozloženie (*Layout*)

V porovnaní s nástrojom Viditeľnosť položky (*Item visiblity*) sa spočiatku zdá byť menej prehľadná. Kombinácia týchto dvoch nástrojov však výrazne uľahčuje reorganizáciu textu a chybného poradia čítania riadkov.

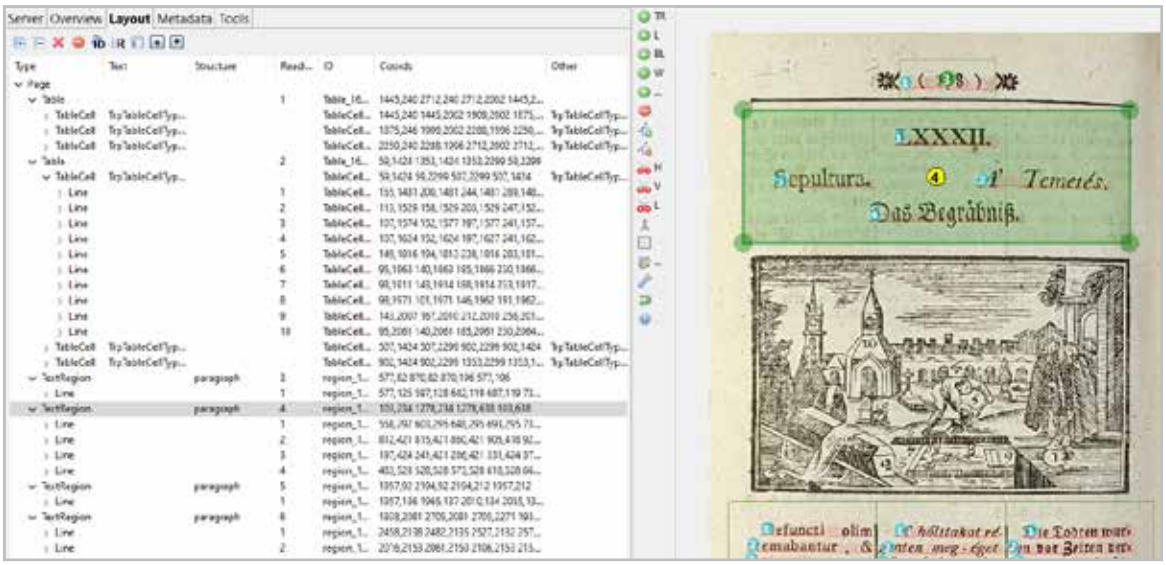
Server | Overview | **Layout** | Metadata | Tools

Type	Text	Structure	Readi...	ID	Coords	Other
Page						
Table			1	Table_16...	1445,240 2712,240 2712,2002 1445,2...	
TableCell	TrpTableCellTyp...			TableCell...	1445,240 1445,2002 1909,2002 1875,...	TrpTableCellTyp...
TableCell	TrpTableCellTyp...			TableCell...	1875,246 1909,2002 2288,1996 2250,...	TrpTableCellTyp...
TableCell	TrpTableCellTyp...			TableCell...	2250,240 2288,1996 2712,2002 2712,...	TrpTableCellTyp...
Table			2	Table_16...	59,1424 1353,1424 1353,2299 59,2299	
TableCell	TrpTableCellTyp...			TableCell...	59,1424 59,2299 507,2299 507,1424	TrpTableCellTyp...
Line			1	TableCell...	155,1481 200,1481 244,1481 289,148...	
Line			2	TableCell...	113,1529 158,1529 203,1529 247,152...	
Line			3	TableCell...	107,1574 152,1577 197,1577 241,157...	
Line			4	TableCell...	107,1624 152,1624 197,1627 241,162...	
Line			5	TableCell...	149,1816 194,1813 238,1816 283,181...	
Line			6	TableCell...	95,1863 140,1863 185,1866 230,1866...	
Line			7	TableCell...	98,1911 143,1914 188,1914 233,1917...	
Line			8	TableCell...	98,1971 101,1971 146,1962 191,1962...	
Line			9	TableCell...	143,2007 167,2010 212,2010 256,201...	
Line			10	TableCell...	95,2061 140,2061 185,2061 230,2064...	
TableCell	TrpTableCellTyp...			TableCell...	507,1424 507,2299 902,2299 902,1424	TrpTableCellTyp...
TableCell	TrpTableCellTyp...			TableCell...	902,1424 902,2299 1353,2299 1353,1...	TrpTableCellTyp...
TextRegion		paragraph	3	region_1...	577,82 870,82 870,196 577,196	
Line			1	region_1...	577,125 597,128 642,119 687,119 73...	
TextRegion		paragraph	4	region_1...	103,234 1278,234 1278,638 103,638	
Line			1	region_1...	558,292 603,295 648,295 693,295 73...	
Line			2	region_1...	812,421 815,421 860,421 905,418 92...	
Line			3	region_1...	197,424 241,421 286,421 331,424 37...	
Line			4	region_1...	483,528 528,528 573,528 618,528 66...	
TextRegion		paragraph	5	region_1...	1957,92 2194,92 2194,212 1957,212	
Line			1	region_1...	1957,136 1965,137 2010,134 2055,13...	
TextRegion		paragraph	6	region_1...	1938,2081 2705,2081 2705,2271 193...	
Line			1	region_1...	2458,2138 2482,2135 2527,2132 257...	
Line			2	region_1...	2016,2153 2061,2150 2106,2153 215...	

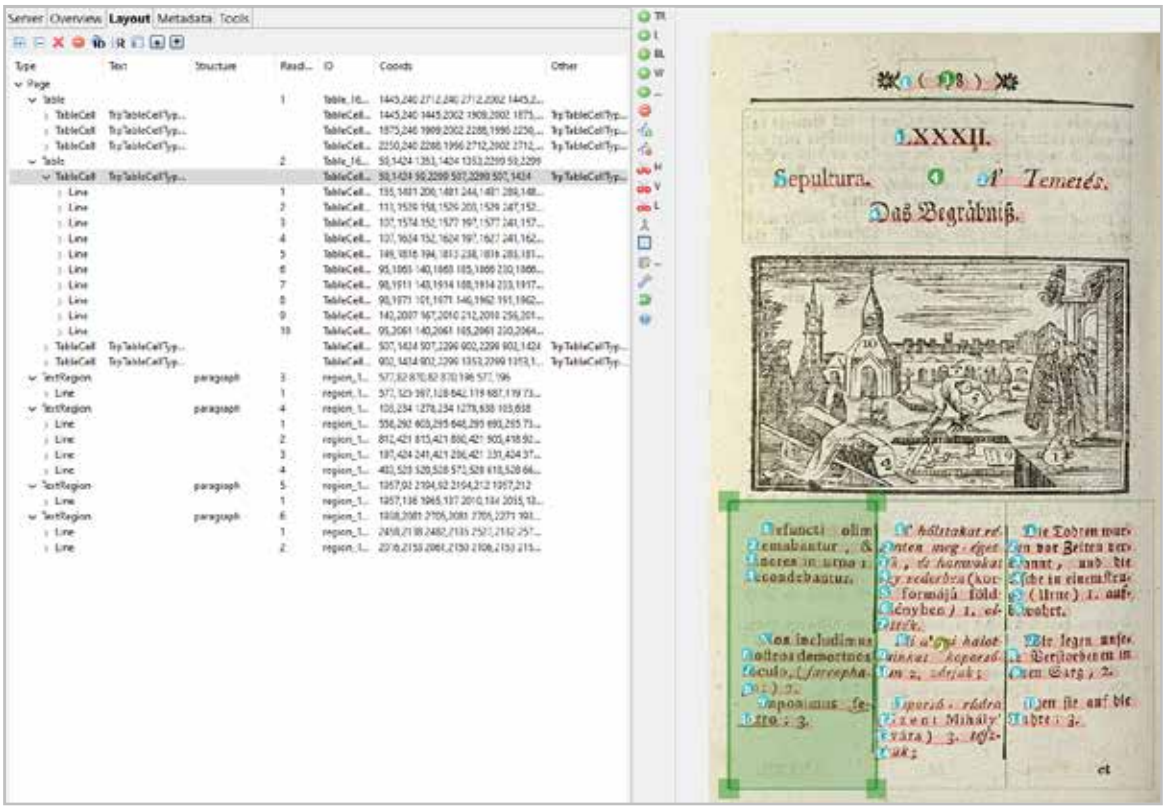
Obrázok 89 Náhľad na štruktúru objektov na záložke Layout

Úprava poradia čítania objektov segmetnácie na záložke *Layout*:

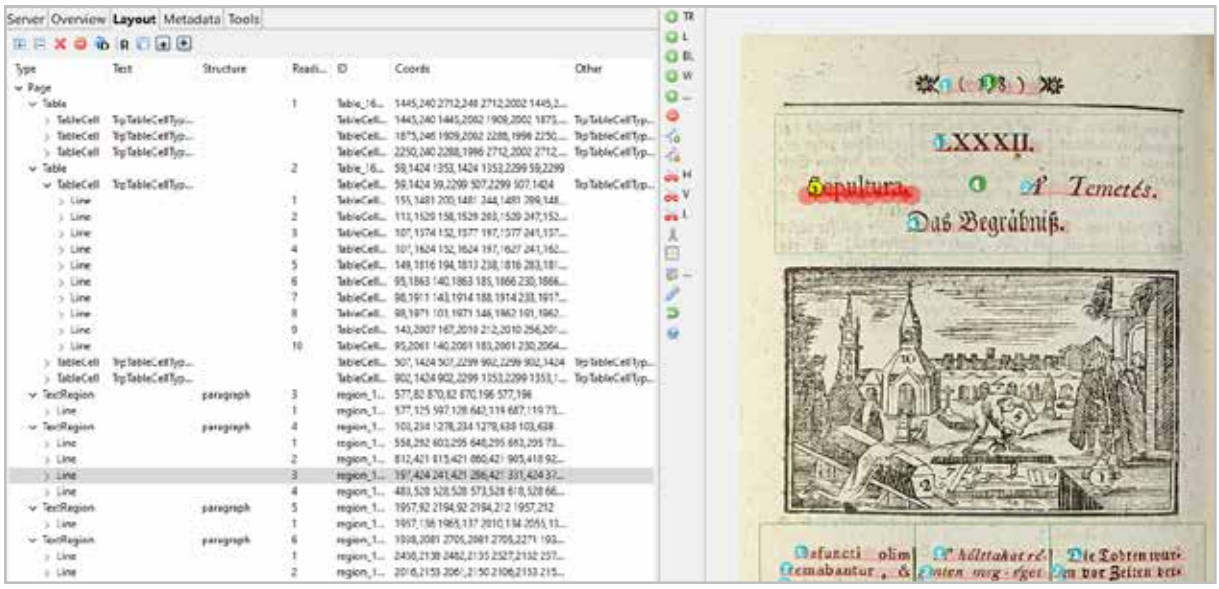
- cez ikonu Viditeľnosť položky (*Item visiblity*) si nastavte zobrazenie číslovania textových rámcov a riadkov (*Show regions reading order* a *Show lines reading order*),
- otvorte záložku *Layout* v menu na ľavej strane okna expert klienta – zobrazí sa zoznam objektov segmentácie s popisom ich typu, štruktúry, poradia čítania, súradnicami na snímke dokumentu a i.,
- kliknutím na zobáčik  pri označení typu objektu, alebo pomocou ikoniek  v záhlaví záložky môžete jednotlivé položky segmentácie rozbaliť alebo minimalizovať,
- kliknutím na objekt na snímke sa tento zvýrazní (sivé podsvietenie),



Obrázok 90 Príklad zobrazenia bloku textu na záložke Layout

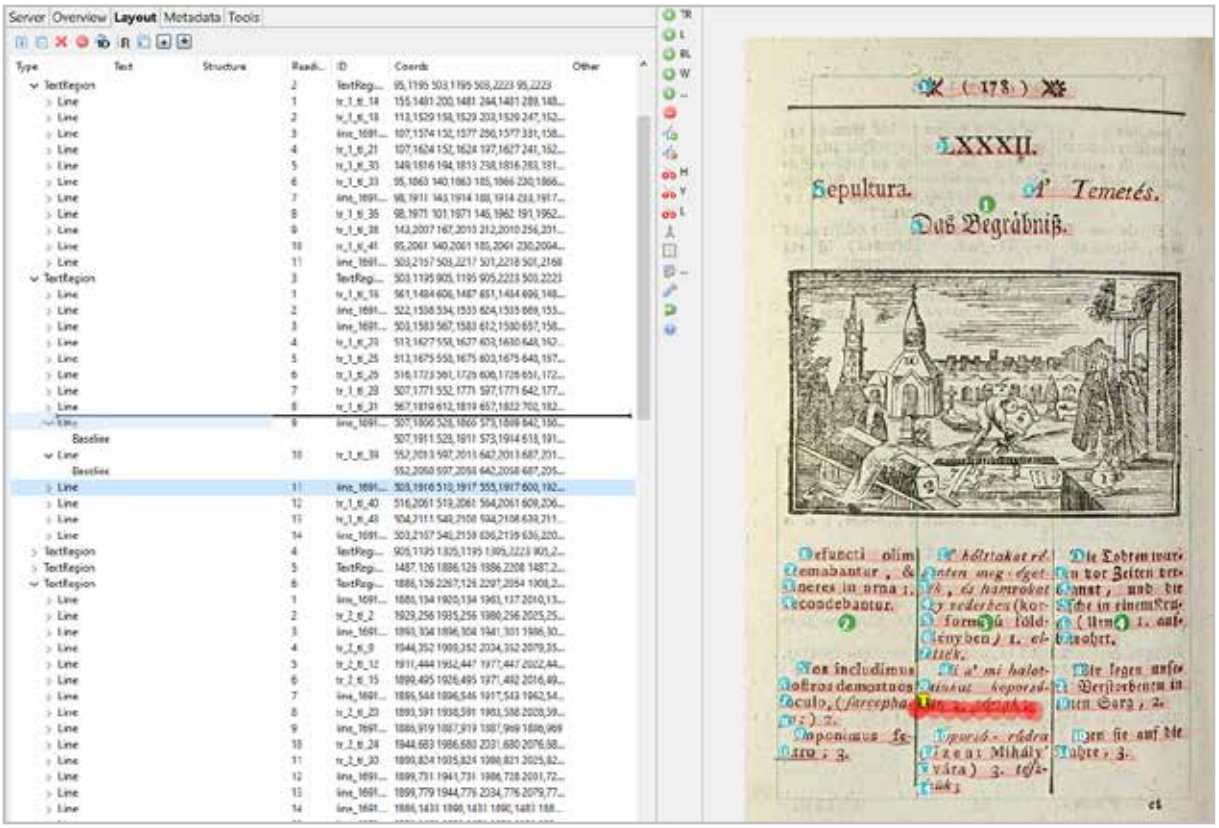


Obrázok 91 Príklad zobrazenia bunky tabuľky na záložke Layout



Obrázok 92 Príklad zobrazenia riadku na záložke Layout

- objekty segmentácie môžete presúvať dvomi spôsobmi:
  - použitím ikoniek  nachádzajúcich sa v hornej časti záložky *Layout*,
  - presúvaním riadkov v textovej štruktúre – vybraný riadok potiahnutím presuniete na požadované miesto (podobne ako presúvate označený text v programe Word), umiestnenie riadka na novom mieste reprezentuje čierna čiara.



Obrázok 93 Presúvanie riadku na záložke Layout

4.2.4.3 Práca so stĺpcami

Dodatočne možno upraviť aj poradie čítania objektov, ktoré majú iné usporiadanie, napr. stĺpce. Program automaticky priraďuje poradie čítania na základe horizontálneho usporiadania riadkov na stránke namiesto toho, aby riadky zoradil podľa stĺpcov. Čiastočne tento problém odstránite nasledovne:

- pomocou ikonky  (*Splits* a *shape with* a *vertical line*) v editore *Canvas* rozdeľte textový rámec podľa usporiadania stĺpcov na snímke,
- keď je každý stĺpec vyčlenený v samostatnom textovom rámci, poradie čítania riadkov sa automaticky aktualizuje.

Na príkladoch nižšie vidieť, že vertikálnym rozdelením stĺpcov došlo aj k rozdeleniu riadkov, ktoré prechádzali cez viacero stĺpcov (napr. riadky č. 6 a č. 21 v prvom stĺpci). Tento krok vo väčšine prípadov vyžaduje následnú kontrolu a korekciu poradia čítania riadkov.

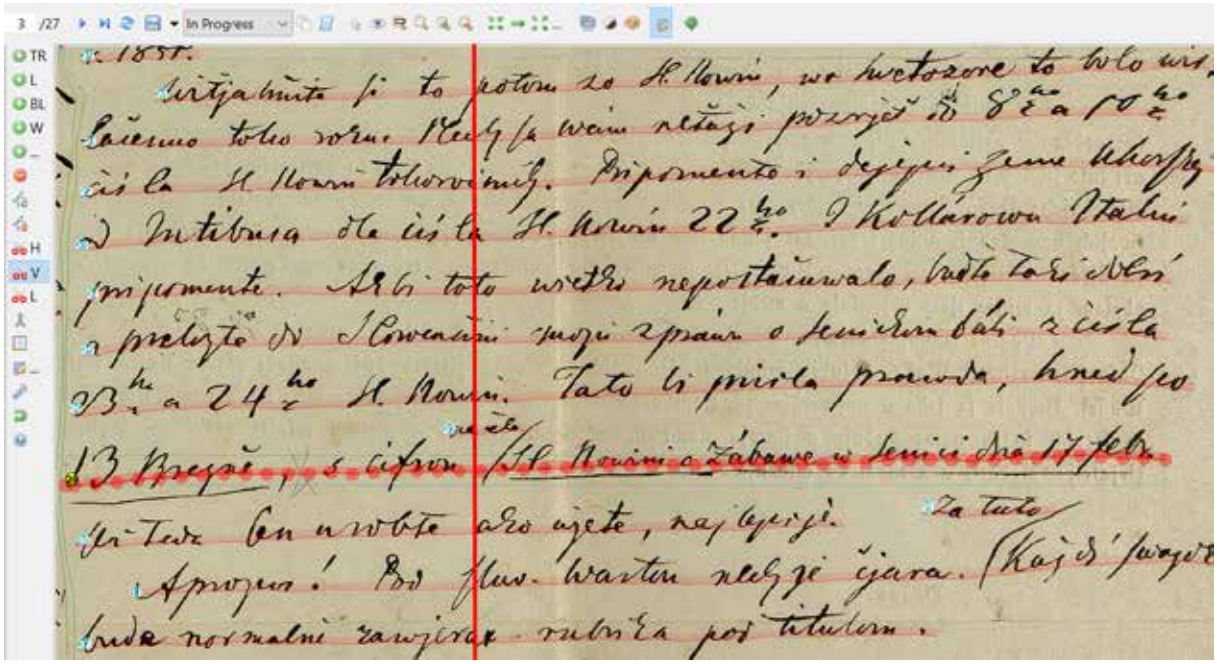
Obrázok 94 Poradie čítania riadkov v stĺpcoch pred vertikálnym rozdelením textového rámca

Obrázok 95 Poradie čítania riadkov v stĺpcoch po vertikálnom rozdelení – v strednom a pravom stĺpci treba poradie korigovať

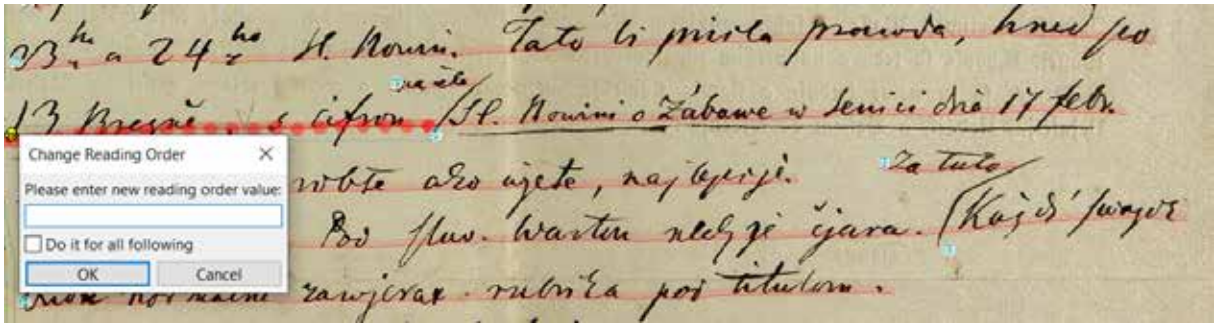
4.2.4.4 Vkladanie medziriadkov

Pri rukopisných textoch sa môžete stretnúť s vloženým textom (vsuvkou), ktorým autor do pôvodného textu vkladá nový obsah. Vložený text vytvára medziriadok, ktorý treba správne včleniť do štruktúry a obsahu dokumentu tak, aby text logicky nasledoval. Na vygenerovanie správneho poradia čítania je potrebné urobiť manuálne úpravy:

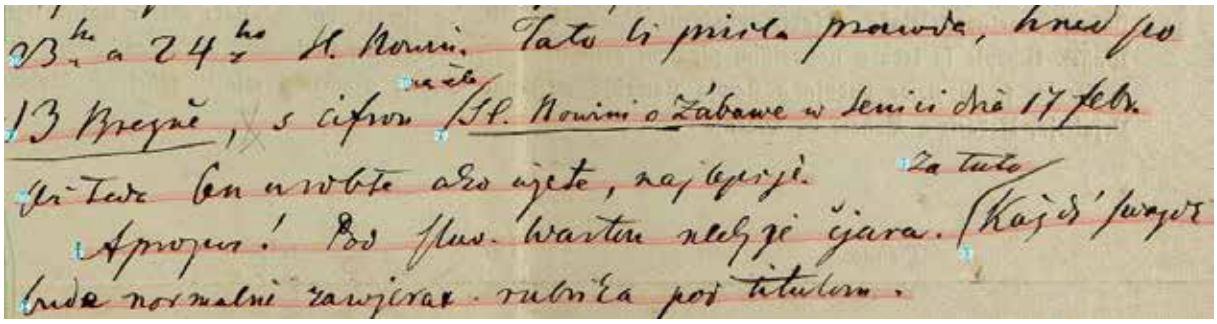
- cez ikonku Viditeľnosť položky (*Item visibility*) si nastavte zobrazenie číslovania riadkov (*Show lines reading order*),
- kliknutím označte riadok nachádzajúci sa pod vloženým textom,
- pomocou ikonky V (*Splits* a *shape with* a *vertical line*) v editore *Canvas* rozdeľte riadok na mieste, kde vložený text obsahovo patrí,
- opravte číslovanie poradia riadkov.



Obrázok 96 Rozdelenie riadka, do ktorého treba vložiť vsunutý text

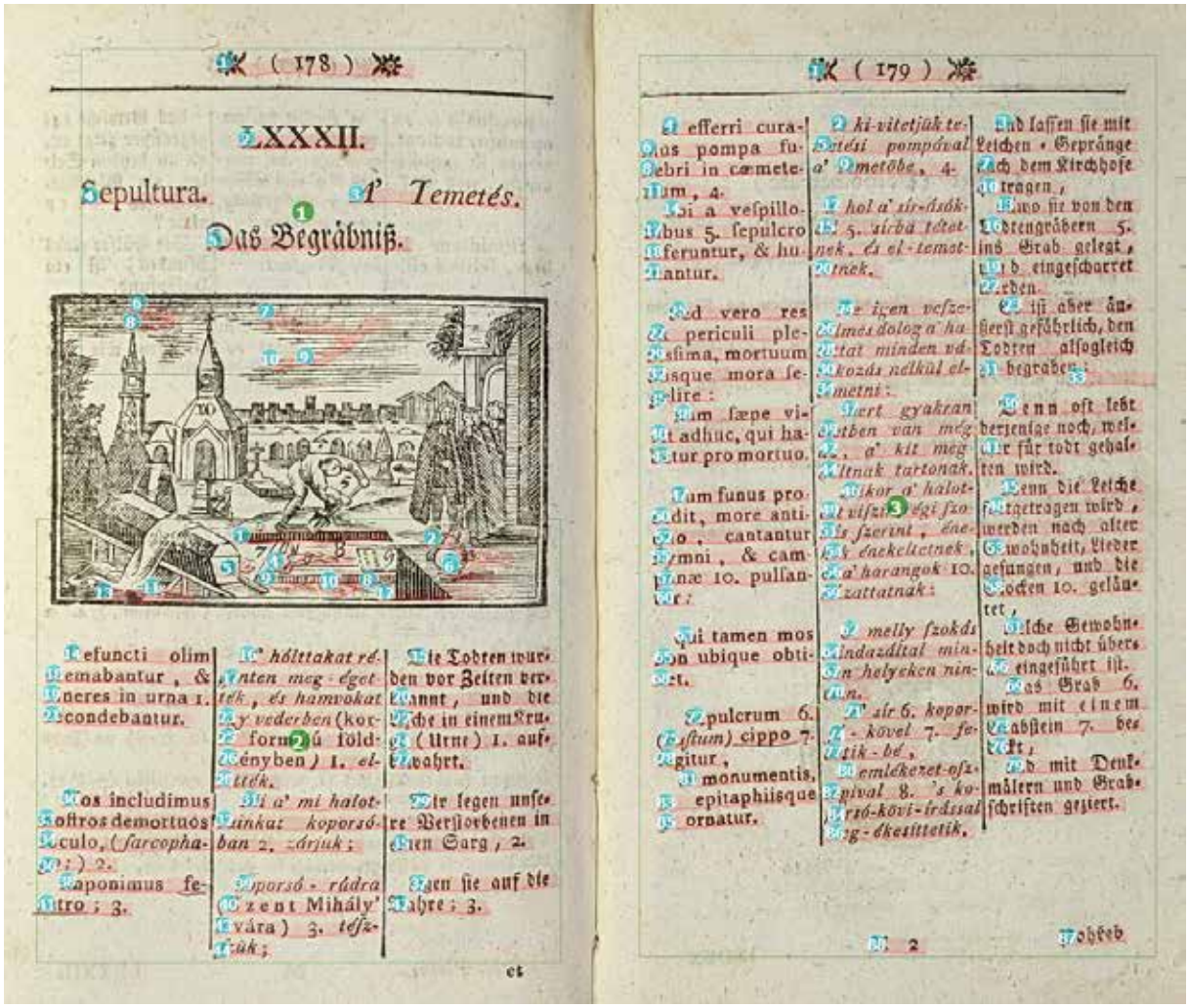


Obrázok 97 Úprava poradia číslovania: riadok č. 28 bude prečíslovaný na č. 27, riadok č. 27 bude mať č. 28, poradie čítania riadku č. 29 je správne



Obrázok 98 Správne poradie číslovania riadkov s vloženým textom po manuálnych úpravách. Na tomto príklade vidieť dva vložené texty s upraveným poradím čítania (riadky č. 28 a č. 32)

Na obrázkoch nižšie uvádzame názorné príklady úpravy stránky po automatickej a manuálnej segmentácii štruktúrovaného textu.



(178)

LXXXII.

Sepultura. Temetés.

Das Begräbniß.

(179)

Obrázok 99 Neuspokojivé výsledky automatickej segmentácie textových a riadkových rámcov pri použití metódy Transkribus LA

Algoritmus na obrázku vyššie automaticky identifikoval tri bloky textu, pričom do dvoch textových rámcov zahrnul aj časť ilustrácie. Pre ľahšiu identifikáciu riadkov by bolo vhodnejšie oddeliť text zapísaný do stĺpcov do samostatných textových rámcov. To je možné urobiť tromi spôsobmi:

1. vytvorením samostatných textových rámcov pre každý stĺpec zvlášť,
2. použitím nástroja na prácu s tabuľkami (viac v kapitole *4.3 Práca s tabuľkami*),
3. dodatočným rozdelením stĺpcov za pomoci funkcie  (*Splits* a *shape with* a *vertical line*) v editore *Canvas* (viac v kapitole *4.2.4.3. Práca so stĺpcami*).

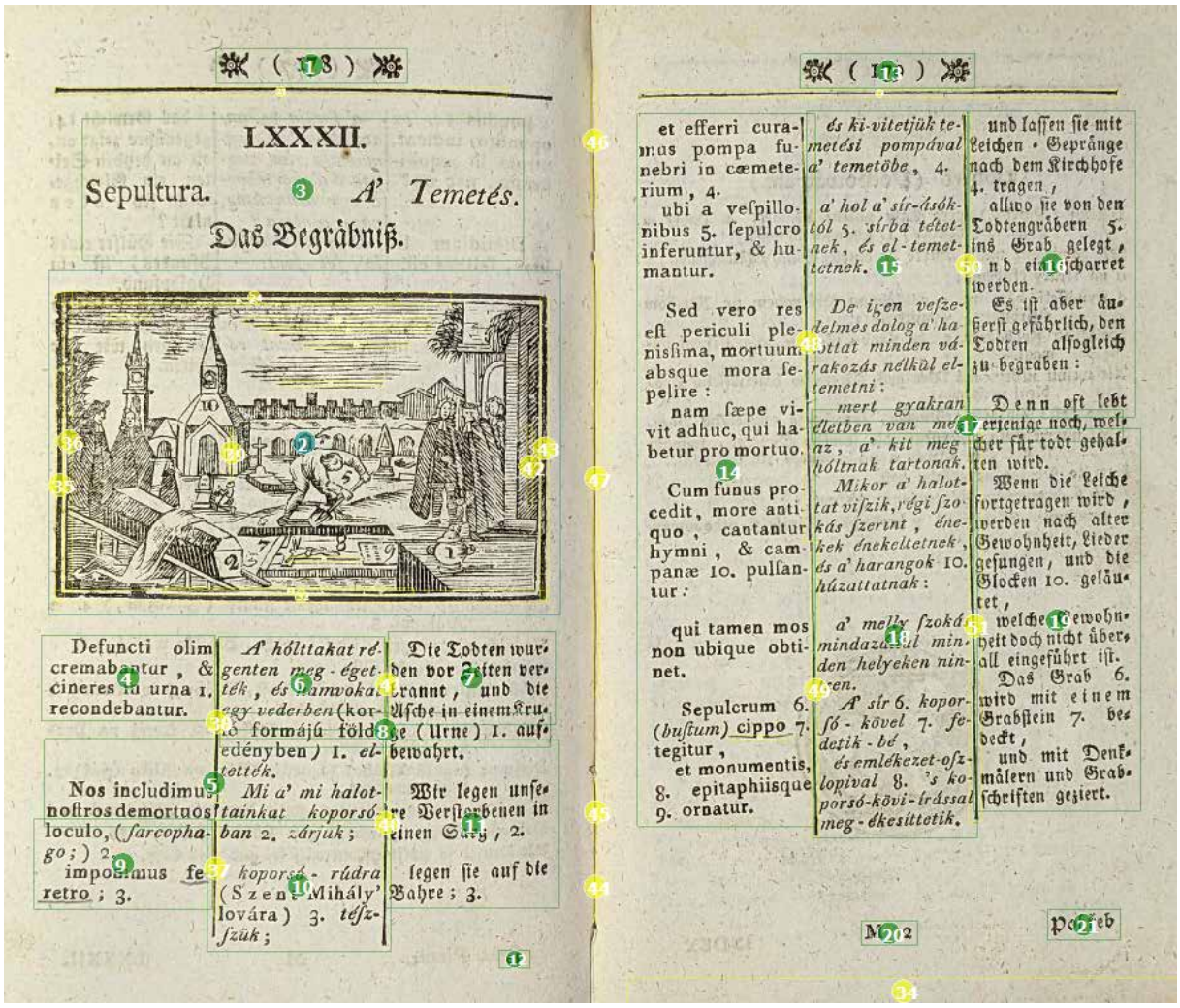
Tým, že nedošlo k správnej segmentácii textových rámcov, nie je správne ani poradie čítania riadkov. Navyše algoritmus detekoval aj riadky v ilustrácii, ktorá zasahuje do dvoch textových rámcov.

Napriek tomu, že text nachádzajúci sa v stĺpcoch nebol pri segmentácii rozdelený do samostatných textových rámcov, softvér automaticky pristúpil k segmentácií textu do stĺpcov. Viditeľných je niekoľko chýb, kde text, ktorý sa nachádza v rôznych stĺpcoch, je spojený do jedného riadku, napr. do dvoch stĺpcov prechádzajú riadok č. 19 na str. 178, riadky č. 18, 26, 28 na str. 179. Text na snímke je mierne naklonený (smeruje zdola nahor), čo má za následok nesprávne poradie čítania riadkov, pretože riadok nachádzajúci sa vyššie má z hľadiska nastavenia algoritmu vyššiu prioritu. Preto sú na str. 179 takmer všetky riadky číslované sprava doľava.

Manuálna korekcia poradia čítania riadkov takto segmentovaného dokumentu by bola časovo náročná (min. 10 – 15 min.). Spočívala by:

- vo vymazaní nesprávne identifikovaných riadkov,
- v doplnení chýbajúcich riadkov,
- v upravení nesprávne vymedzených riadkov, napr. riadok č. 9 v strednom stĺpci na str. 179,
- v rozdelení spojených riadkov, ktoré sa majú nachádzať v rôznych textových rámcoch (stĺpcoch),
- v rozdelení textových rámcov do stĺpcov s použitím funkcie  (*Splits* a *shape with vertical line*) v editore *Canvas*, prípadne aj oddelením ostatných častí textu, napr. číslovanie, kustódy s použitím funkcie  (*Splits* a *shape with* a *horizontal line*) v editore *Canvas*,
- v usporiadaní textových rámcov do správneho poradia,
- v kontrole poradia čítania riadkov v textových rámcoch a presune nesprávneho poradia riadkov na záložke *Layout*.

Nesprávne usporiadanie textu nemá vplyv na vytváranie modelu, pretože softvér sa učí čítať jednotlivé znaky bez ohľadu na logické usporiadanie textu. Taktiež nemá vplyv na následnú transkripciu dokumentu. Sťažuje však transkripciu nevyhnutného počtu strán potrebných na trénovanie modelu a zároveň komplikuje percepciu prepísaného dokumentu.



Obrázok 100 Neuspokojivé výsledky automatickej segmentácie textových a riadkových rámcov s použitím metódy Printed Block Detection

Algoritmus na obrázku vyššie automaticky identifikoval textové rámce. Segmentáciou bolo vytvorených niekoľko rámcov, pričom vo viacerých prípadoch bol správne oddelený text nachádzajúci sa v stĺpcoch. Ako samostatný textový rámec bola identifikovaná aj ilustrácia. Do segmentácie boli zároveň zahrnuté aj ozdoby tlače pri paginácii, ktoré by z pohľadu trénovania modelu mohli pôsobiť rušivo.

V niektorých prípadoch nedošlo k správnemu oddeleniu textu v stĺpcoch, čo spôsobilo určenie nesprávneho poradia čítania textových rámcov a poradie čítania riadkov.

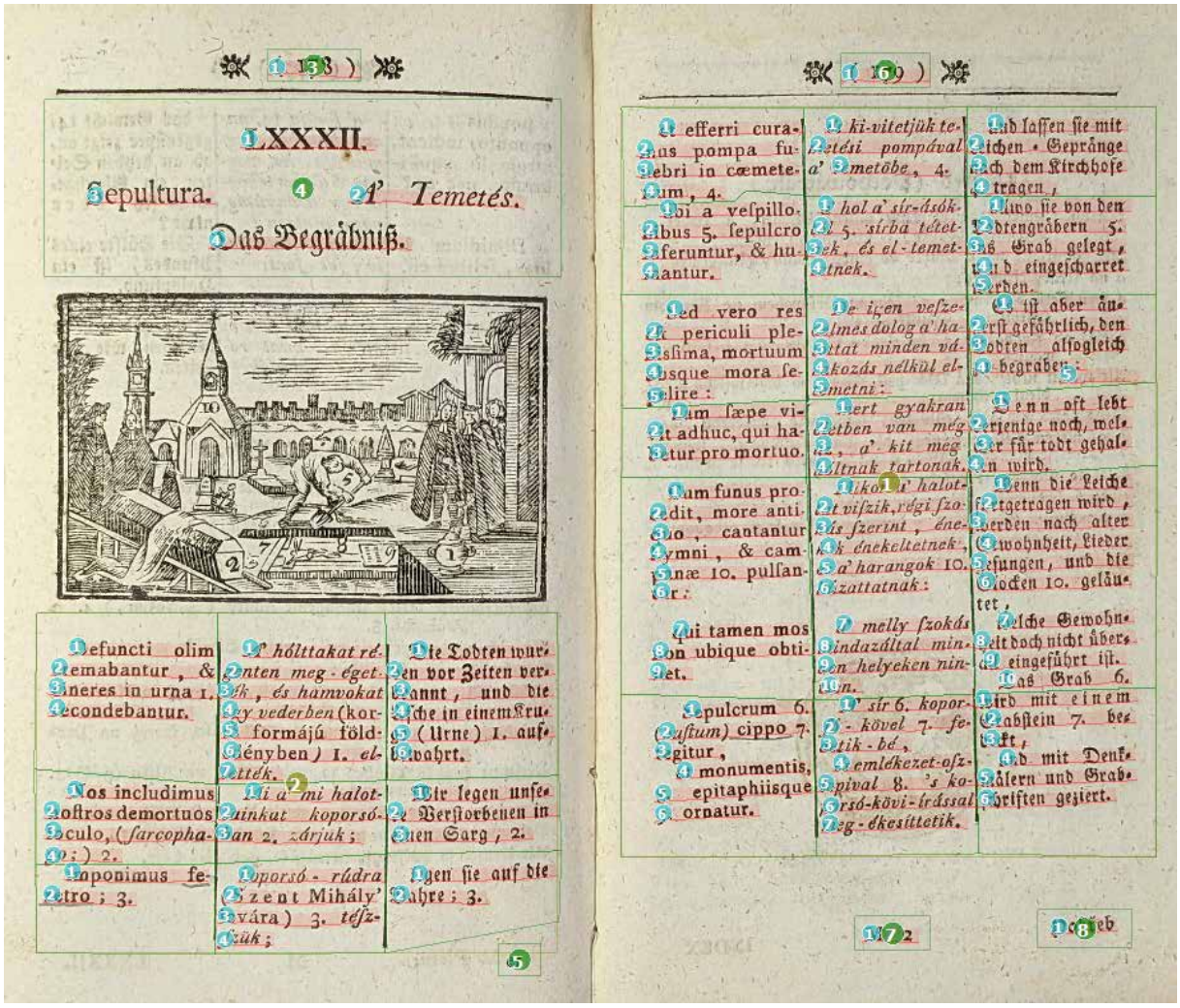
Aj napriek detekcii samostatných textových rámcov v stĺpcoch je viditeľných niekoľko chýb, kde text, ktorý sa nachádza v rozličných stĺpcoch, je spojený do jedného riadku, napr. dvojitý výskyt riadkov č. 5, 8, 11 a 15 v bloku 18 v strednom stĺpci na str. 179.

Manuálna korekcia poradia čítania riadkov takto segmentovaného dokumentu by bola časovo náročná (min. 10 min.). Spočívala by:

- vo vymazaní nesprávne identifikovaných riadkov,
- v doplnení chýbajúcich riadkov,
- v upravení nesprávne vymedzených riadkov, napr. druhý výskyt riadku č. 2 v strednom stĺpci na str. 179,

- v rozdelení spojených riadkov, ktoré sa majú nachádzať v rôznych textových rámcoch (stĺpcoch),
- v rozdelení textových rámcov do stĺpcov s použitím funkcie V (*Splits* a *shape with vertical line*) v editore *Canvas*,
- v usporiadaní textových rámcov do správneho poradia,
- v kontrole poradia čítania riadkov v textových rámcoch a presune nesprávneho poradia riadkov na záložke *Layout*.

Nesprávne usporiadanie textu nemá vplyv na vytváranie modelu, pretože softvér sa učí čítať jednotlivé znaky bez ohľadu na logické usporiadanie textu. Taktiež nemá vplyv na následnú transkripciu dokumentu. Sťažuje však transkripciu nevyhnutného počtu strán potrebných na trénovanie modelu a zároveň komplikuje percepciu prepísaného dokumentu.



Obrázok 101 Uspokojivé výsledky manuálnej segmentácie textových rámcov a automatickej segmentácie riadkov

Bloky textu na obrázku vyššie sú manuálne rozčlenené do viacerých textových rámcov, ktoré označujú jednotlivé časti textu. Nie sú však logicky správne usporiadané. Na oddelenie textu buniek bol použitý nástroj na segmentáciu tabuliek (viac v kapitole *4.3 Segmentácia tabuliek*).

Použitím funkcie *Split lines on regions* pri nastavení automatickej segmentácie riadkov nedošlo k spojeniu riadkov prechádzajúcich medzi jednotlivých stĺpcami textu.

V segmentácii riadkov je viditeľných niekoľko chýb, napr. neoznačená kustóda na str. 178, neoznačený riadok v piatej bunke pravého stĺpca na str. 179, nedotiahnutá základná čiara riadku č. 3 v prvej bunke stredného stĺpca na str. 179, nesprávne detekovaný riadok č. 5 v tretej bunke pravého stĺpca na str. 179 a i.

Manuálna korekcia chýb takto segmentovaného dokumentu trvá približne dve minúty.

Obrázok 102 Upravené číslovanie poradia čítania objektov segmentácie

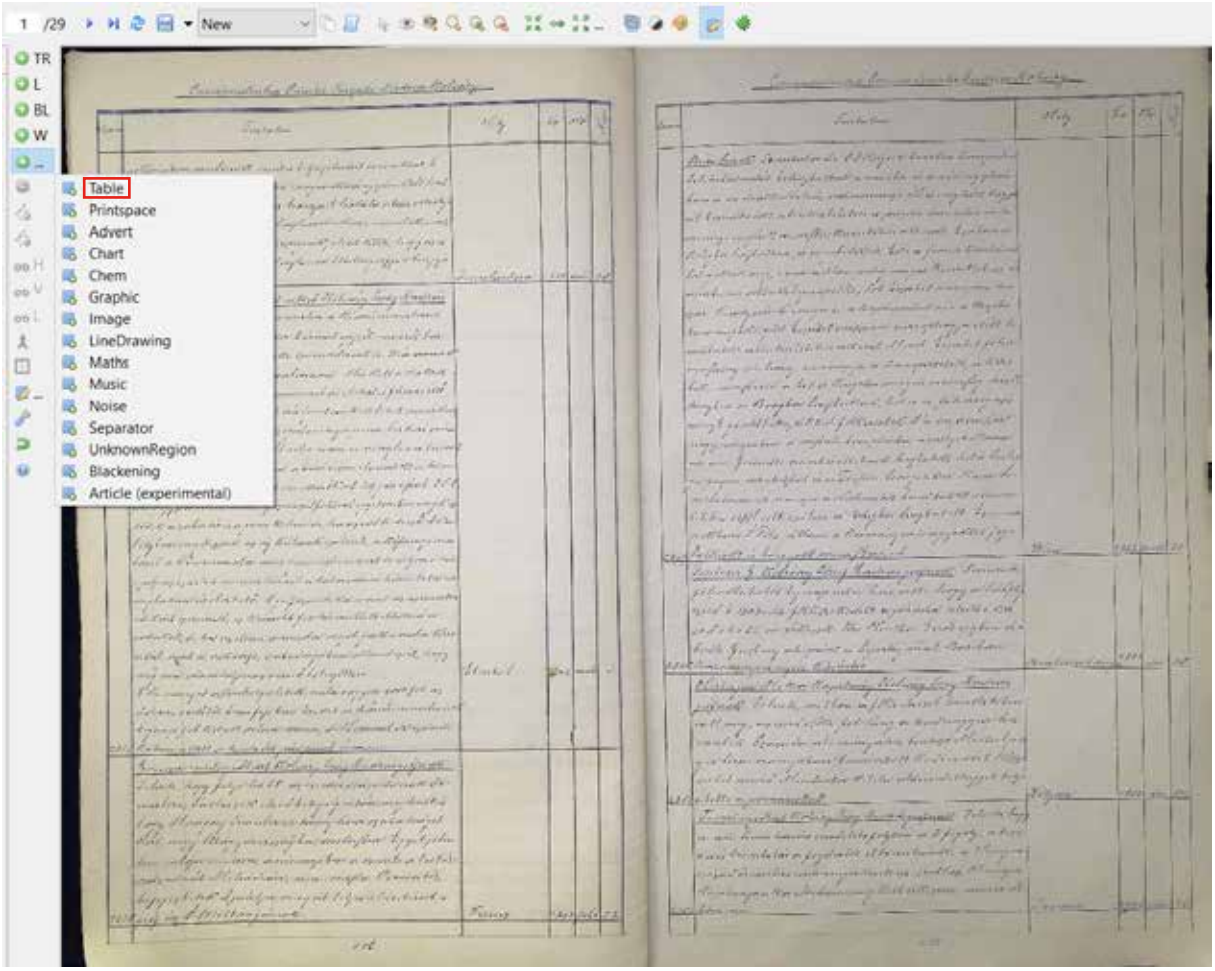
4.3 Segmentácia tabuliek

Segmentácia tabuliek v Transkribus expert klientovi je poloautomatický proces. Najskôr treba vytvoriť štruktúru tabuľky a následne spustiť automatickú segmentáciu riadkov. Segmentovanie tlačených a ručne kreslených tabuliek umožňuje nástroj **Tabuľka** (*Table*) v editore *Canvas*. Vďaka nemu si manuálne vytvoríte vonkajšie hranice tabuľky. Takto zadefinovaná oblasť tabuľky následne zjednodušuje a zefektívňuje využívanie ostatných funkcií editora *Canvas* na tvorbu vnútornej štruktúry, t. j. rozdelenie textu do stĺpcov a riadkov.

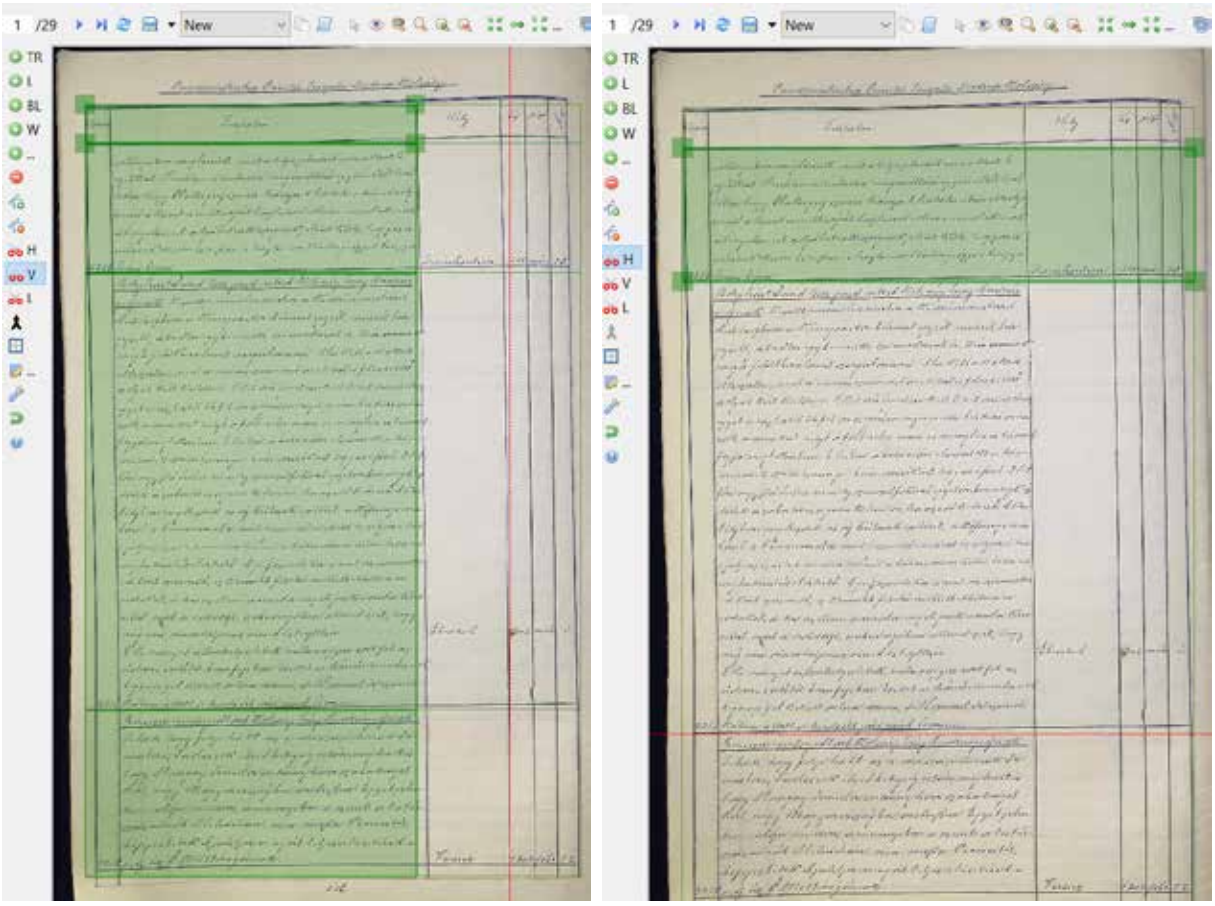
Pre manuálne vytvorenie tabuľky postupujte nasledovne:

- v editore *Canvas* kliknite na ikonku ... (*Add other Item...*),
- zo zoznamu nástrojov vyberte voľbu Tabuľka (*Table*),
- na snímke dokumentu označte celú oblasť tabuľky,

- pomocou funkcie pre horizontálne delenie H (*Split* a *shape with horizontal line*) v editore *Canvas* rozdeľte tabuľku na riadky – kliknite na všetky čiary, ktoré definujú spodné línie buniek tabuľky,
- pomocou funkcie pre vertikálne delenie V (*Split* a *shape with vertical line*) v editore *Canvas* vytvorte v tabuľke stĺpce – kliknite na všetky čiary, ktoré definujú bočné línie buniek tabuľky.

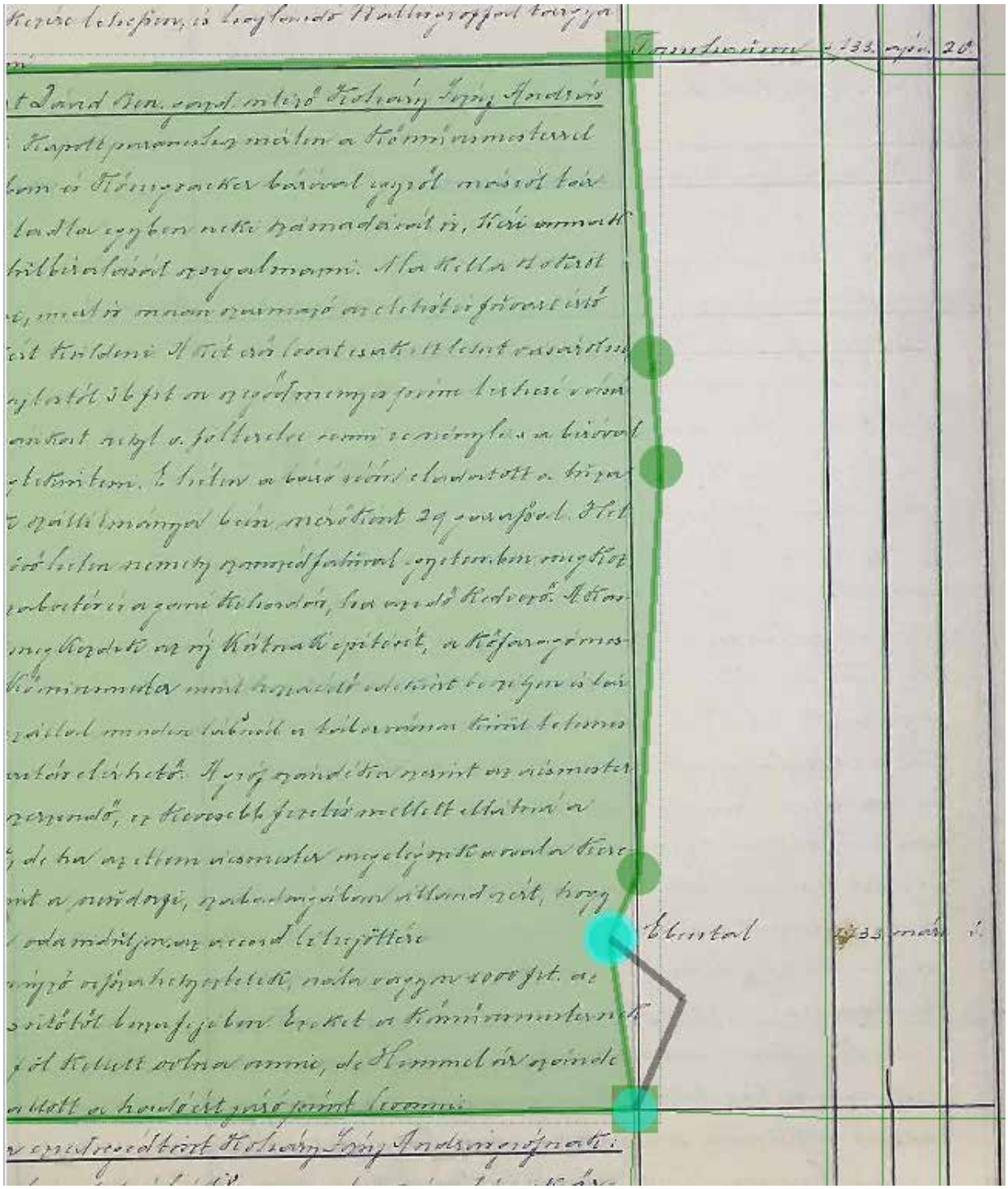


Obrázok 103 Výber funkcie na segmentáciu tabuliek



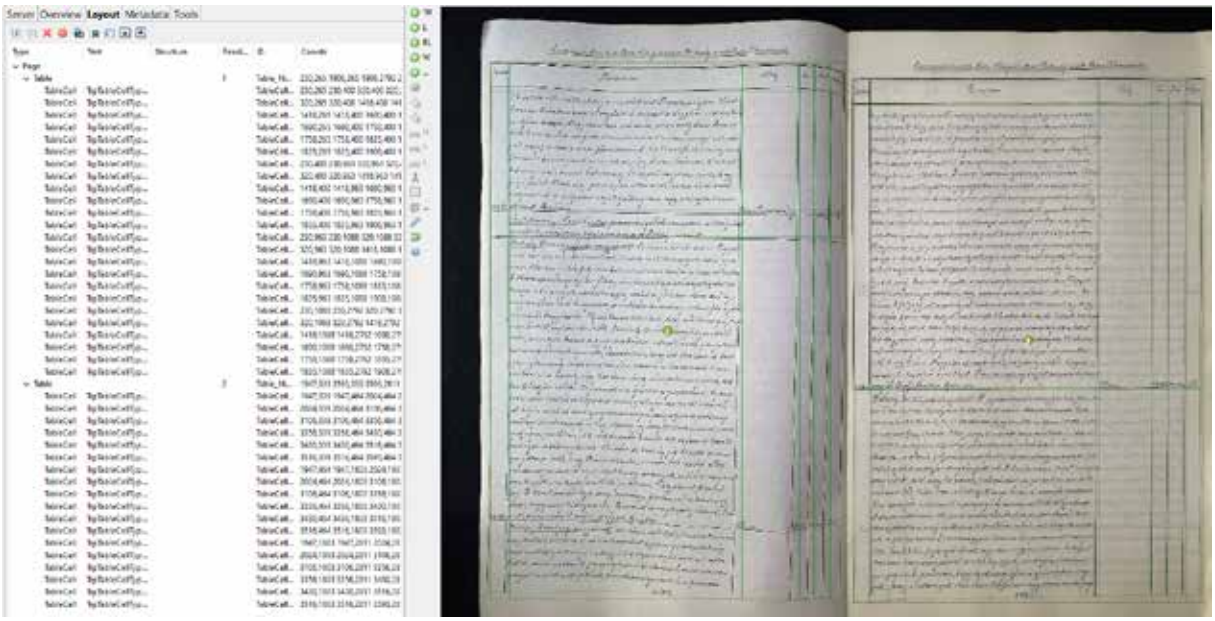
Obrázok 104 Horizontálne a vertikálne členenie tabuľky

Po rozdelení tabuľky na bunky je zvyčajne potrebné manuálne korigovať čiary, ktoré tvoria tvar jednotlivých buniek. Môžete tak urobiť posúvaním bodov/vrcholov oblasti bunky alebo pridávaním ďalších bodov pomocou funkcie  (*Add point to selected shape*) na vytvorenie špecifických polygónov kopírujúcich text umiestnený v bunke.



Obrázok 105 Detail – korekcia buniek tabuľky pridávaním kontrolných bodov

Segmentovaná tabuľka predstavuje jeden blok textu automaticky označený ako *Table*. V záložke *Layout* môžete skontrolovať poradie čítania jednotlivých buniek (*TableCell*). Poradie čítania jednotlivých buniek je automaticky nastavené po riadkoch od ľavého horného rohu k pravému dolnému rohu tabuľky.



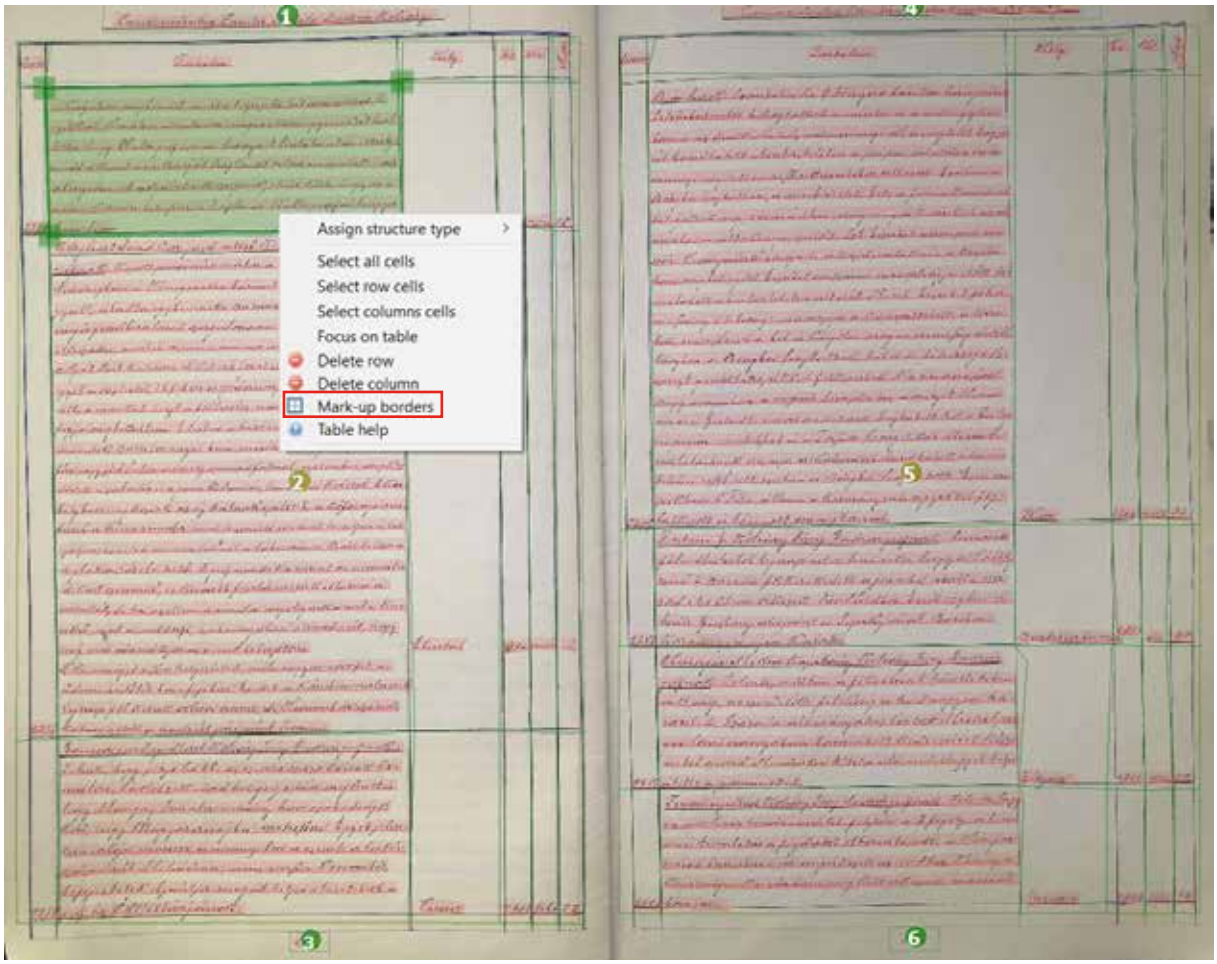
Obrázok 106 Označenie blokov textu (Table) a bunky (TableCell) na záložke Layout

Po ukončení segmentácie tabuľky môžete pristúpiť k automatickej alebo manuálnej segmentácii riadkov a základných čiar textu (pozri kapitolu *4.1 Spôsoby segmentácie*) a kontrole poradia čítania riadkov (pozri kapitolu *4.2.4 Kontrola a úprava poradia čítania textových a riadkových rámcov*).

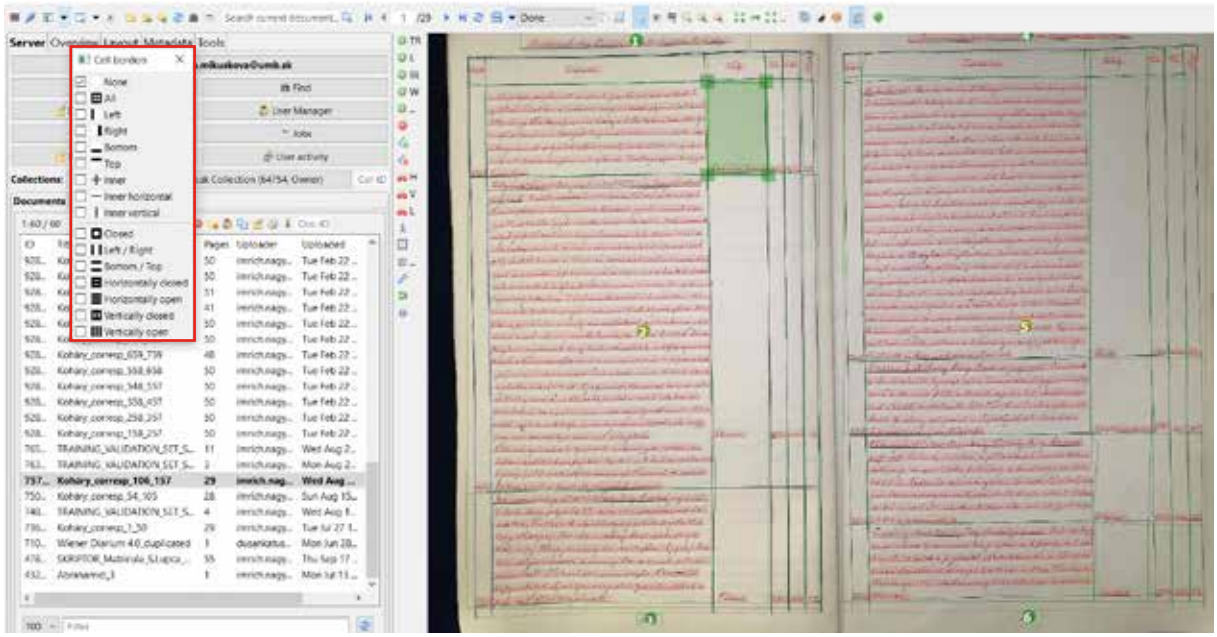
4.3.1 Označenie hraníc bunky

Tabuľka, ktorú ste si vytvorili, slúži najmä pre potreby transkripcie dokumentu. Pre potreby ďalšieho spracovania dokumentu môžete tabuľku graficky upraviť vložením hraníc, ktoré vymedzujú a oddeľujú bunky:

- označte si bunku, ktorej hranice chcete vymedziť,
- stlačte pravé tlačidlo myši,
- v dialógovom okne označte voľbu *Mark-up borders*,
- otvorí sa ďalšie dialógové okno s ponukou, vyberte hranicu, ktorú chcete pri danej bunke vymedziť (podobne ako práca s označením buniek v programe Excel).



Obrázok 107 Dialógové okno s ponukou na prácu s bunkami

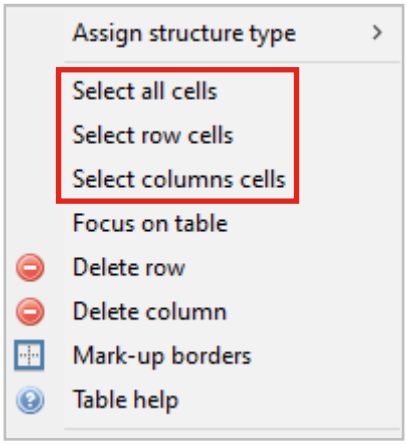


Obrázok 108 Dialógové okno na vyznačenie hraníc bunky

Hranice v tabuľke môžete vyznačovať aj hromadne:

- pomocou voľby Vybrať všetky bunky (*Select all cells*) označte všetky bunky tabuľky,
- pomocou voľby Vybrať všetky bunky v riadku (*Select row cells*) označte všetky bunky v riadku,
- pomocou voľby Vybrať všetky bunky v stĺpci (*Select columns cells*) označte všetky bunky v stĺpci.

Po výbere požadovanej voľby na hromadné označenie buniek v dialógovom okne znovu kliknite na možnosť Označiť hranice (*Mark-up borders*).



Obrázok 109 Voľby na označenie viacerých buniek v tabuľke

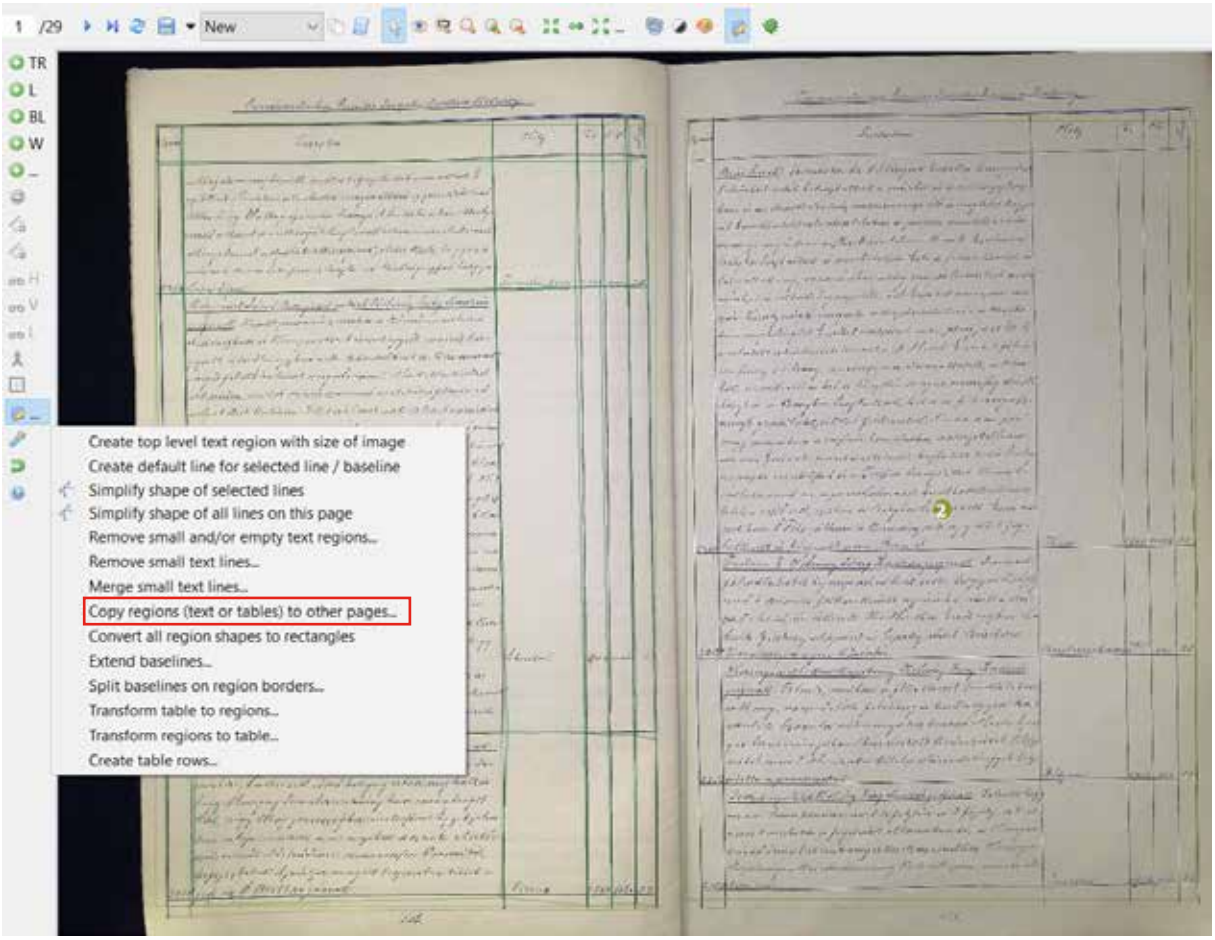
Ak potrebujete ohraničiť bunky, ktoré sa na snímke nenachádzajú vedľa seba, stlačte kláves CTRL a postupným klikaním kurzorom vyznačte príslušné bunky. Následne vyberte voľbu pre označenie hraníc (*Mark-up borders*).

4.3.2 Kopírovanie tabuliek

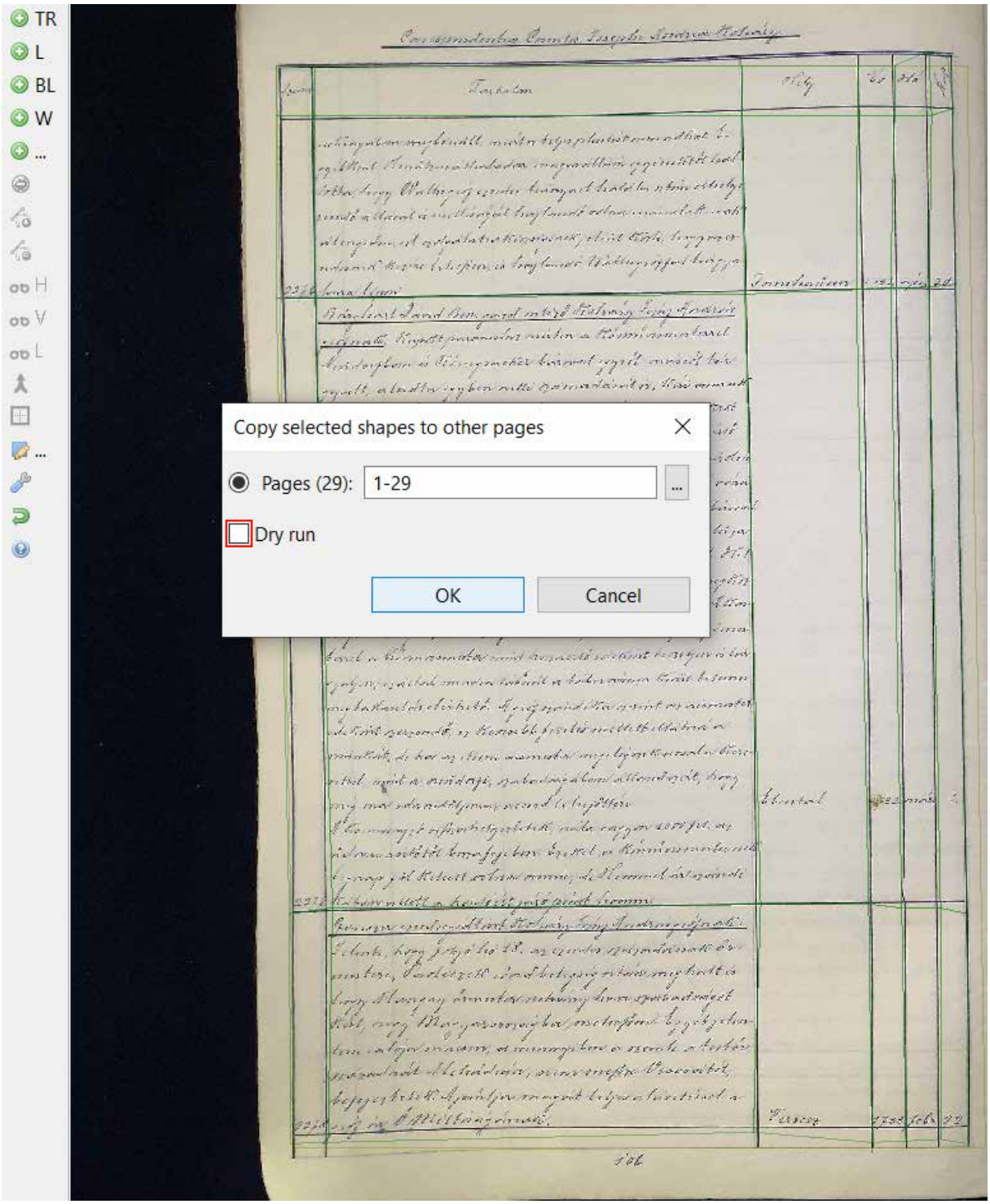
V prípade, že tabuľka sa v rovnakej alebo podobnej štruktúre nachádza na viacerých stranách dokumentu, štruktúru tabuľky stačí vytvoriť len pri prvom výskyte tabuľky a následne ju kopírovať aj na iné snímky dokumentu pomocou nástrojov *nomacs*. Postupujte nasledovne:

- v editore *Canvas* kliknite na ikonu ... Ďalšie nástroje segmentácie (*Other segmentation tools*),
- kliknite na voľbu Kopírovať rámce (*Copy regions (text or tables) to other pages…*),
- v dialógovom okne zapíšte rozsah strán, na ktoré sa má štruktúra tabuľky skopírovať a voľbu potvrďte kliknutím na tlačidlo OK,

POZOR! políčko *Dry run* nesmie byť zaškrtnuté.



Obrázok 110 Voľba na kopírovanie textových rámcov



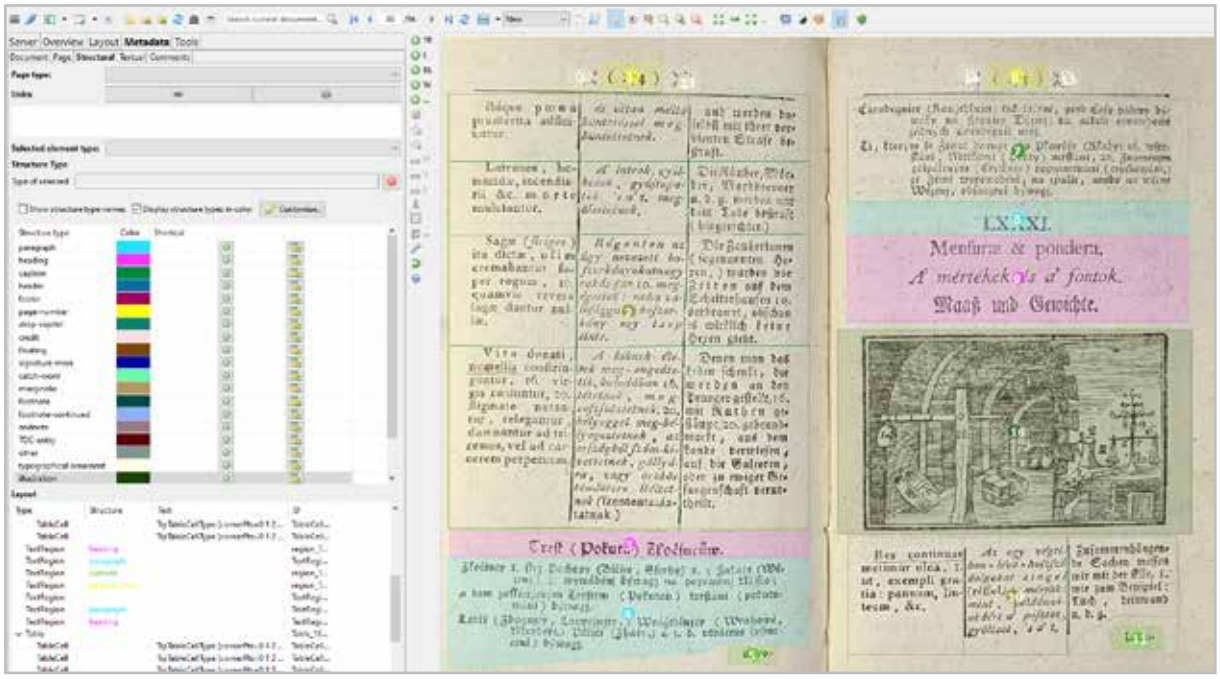
Obrázok 111 Odstránenie voľby Dry run

Je možné, že pozíciu skopírovanej tabuľky/buniek bude potrebné upraviť. Ak chcete upraviť pozíciu celej tabuľky:

- tabuľku označte,
- na klávesnici stlačte CTRL,
- premiestnite tabuľku.

4.3.3 Štrukturálne metadáta – tagovanie štruktúry dokumentu

Transkribus expert klient obsahuje funkcie na značkovanie prvkov štruktúry dokumentu, t. j. popisu obsahu textových rámcov, obsahu bunky a pod. Tieto funkcie na nachádzajú na záložke Metadáta (*Metadata*) v ľavom hornom menu klienta, v podzáložke Štrukturálne tagy (*Structural*) (viac o štrukturálnych tagoch v kapitole *7.2 Štrukturálne metadáta*).



Obrázok 112 Príklad použitia štrukturálnych metadát

Na obrázku sú určené prvky štruktúry takto:

1 a 6 – číslo strany
3 a 9 – názov
5 a 11 – kustóda
8 – číslo kapitoly
12, 13, 14 a 15 – typografická ozdoba
16 – ilustrácia

Transkribus expert klient umožňuje vytrénovať modely tak, aby automaticky rozpoznali štruktúru dokumentu. Cvičné dáta potrebné na tento proces vytvoríte pridaním štrukturálnych značiek pre všetky alebo vybrané prvky dokumentu. Netreba označovať všetky položky, stačí sa zamerať na označenie sekcií, ktoré sú zaujímavé pre ďalšie spracovanie dokumentu alebo pre jeho čitateľov.

5 Tvorba modelu automatickej transkripcie

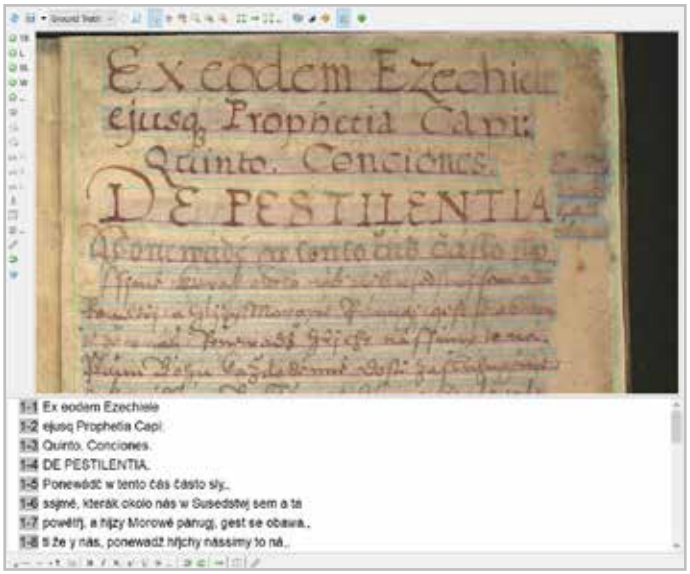
Transkribus expert klient umožňuje trénovaním vytvoriť vlastný model na rozpoznávanie rukopisných alebo tlačených textov, ktorý sa potom použije na automatickú transkripciu celej zbierky dokumentov. Využíva podoblasť umelej inteligencie – nástroj strojového učenia *PyLaia*.

Proces tvorby modelu PyLaia na jeden alebo viacero rukopisov v prostredí expert klienta zahŕňa prepis dokumentu (prípravu vzorky *Ground Truth*), trénovanie modelu PyLaia, vyhodnotenie úspešnosti modelu a jeho zdokonaľovanie

5.1 Prepis dokumentu (príprava vzorky *Ground Truth*)

Prepisy na platforme Transkribus môžu byť použité na trénovanie modelu PyLaia a tiež ako základ pre vytvorenie knižnej alebo digitálnej pramennej edície. Na trénovanie modelu postačuje jednoduchý prepis. Účinnosť modelu závisí od kvality trénovaného materiálu (manuálnej transkripcie), kvality digitalizátov a ich paleografickej náročnosti. Existujú aj pokročilé možnosti prepisu pre prípravu digitálnej edície. Obsahujú napríklad úpravu poradia textu, použitie historických znakov, pridávanie značiek (tagov), metadát a rozpisovanie skratiek.

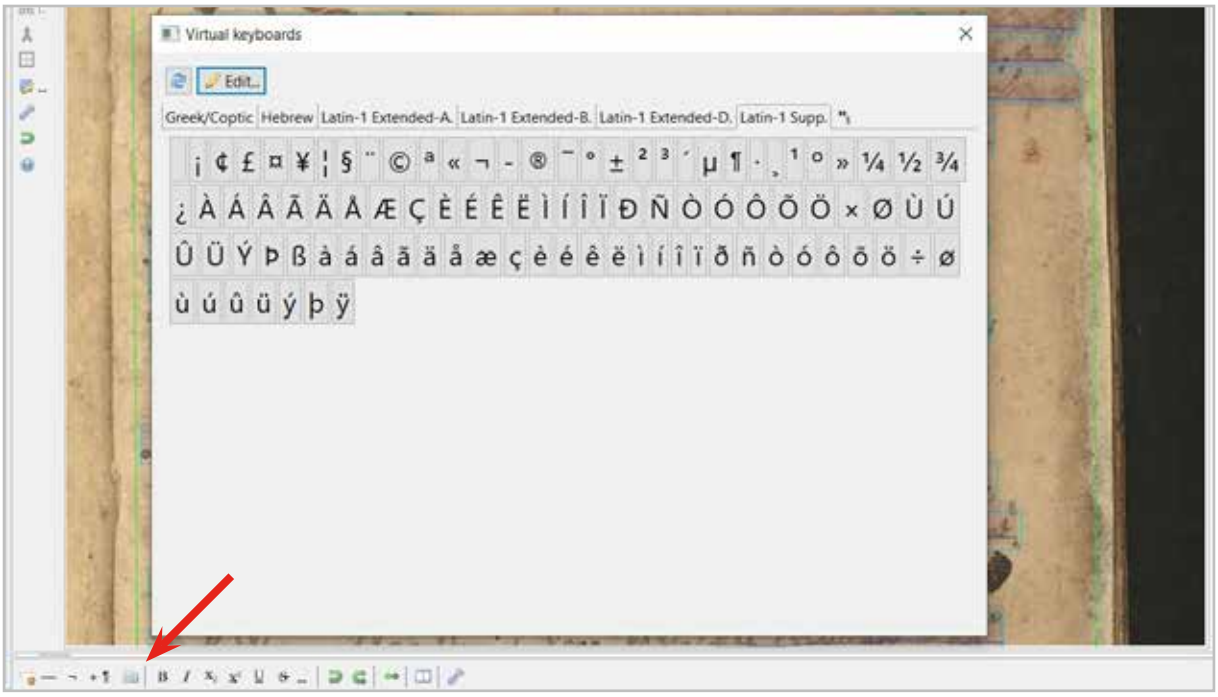
Jednoduchý prepis na trénovanie modelu PyLaia
Po segmentácii dokumentu vyberte možnosť *Transcription*, ktorá sa nachádza v ponuke po stlačení ikony *Profiles* v hlavnom menu. Pod digitalizovaným textom sa zobrazí pole textového editora. Pre každý (základný) riadok na obrázku existuje zodpovedajúci riadok v textovom editore. Zachovajte rovnaké poradie (číslovanie) riadkov v textovom editore aj v segmentovanom dokumente. Prepíšte text podľa zdrojového dokumentu. Dokument môže prepisovať viac používateľov, ale nemali by spracovávať rovnakú stranu dokumentu súčasne.



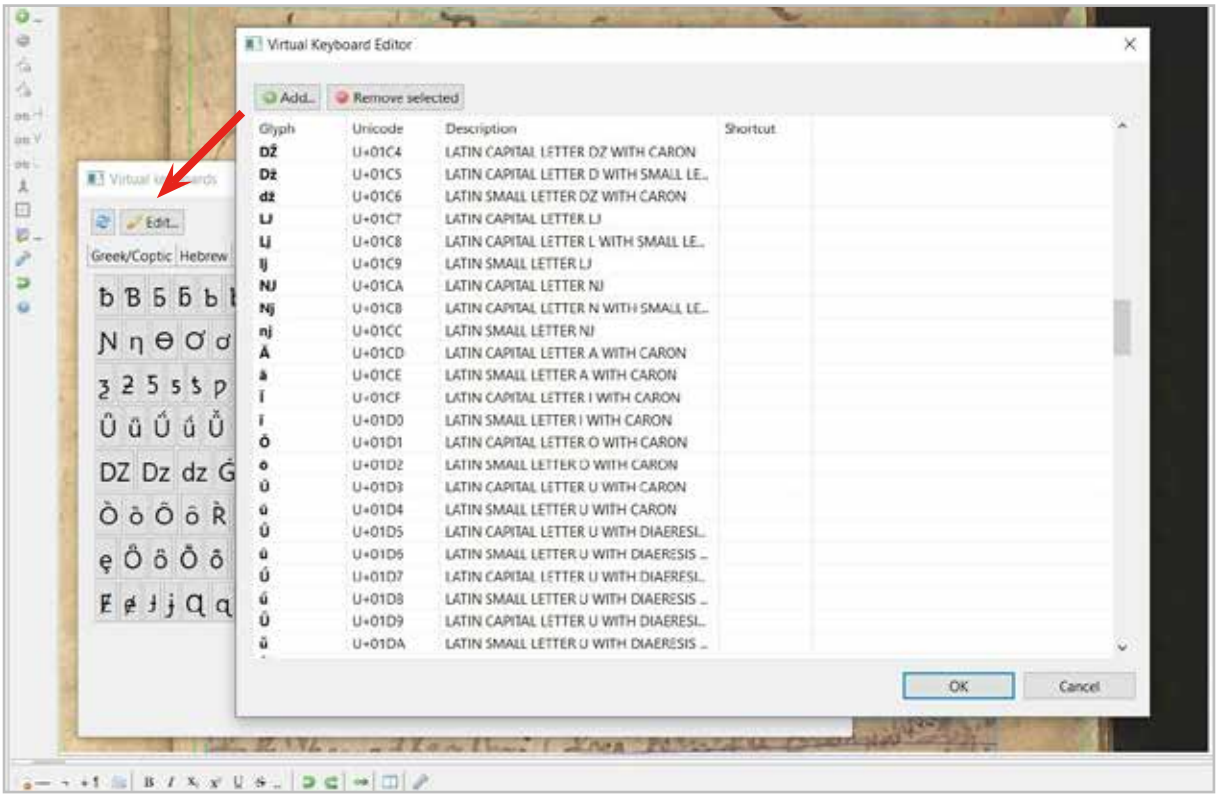
Obrázok 113 Prepis segmentovaného dokumentu

Transkripcia a virtuálna klávesnica
Prepis, ktorý bude slúžiť ako základ pre vedeckú edíciu, by mal používateľovi poskytnúť viac kontextových údajov ako jednoduchý prepis. V tomto prípade zohráva dôležitú úlohu nielen strojové čítanie, ale aj vlastné čítanie užívateľa.

Počas prepisu môžete pridať špeciálne znaky a symboly Unicode (štandardizovaná schéma pre písané jazyky) s použitím **virtuálnych klávesníc** (*Virtual keyboards*) v poli textového editora. V prípade súbežnej práce viacerých osôb s rôznymi verziami Transkribu je potrebné jednotné používanie znakov z konkrétnej klávesnice. Znaky z rozdielnych klávesníc sa totiž môžu prejaviť v zvýšenej chybovosti počas následného trénovania modelov. Tlačidlom *Edit...* je možné pridať klávesové skratky pre často používané znaky a pridať nové znaky Unicode.



Obrázok 114 Virtuálna klávesnica Latin-1 Supp.



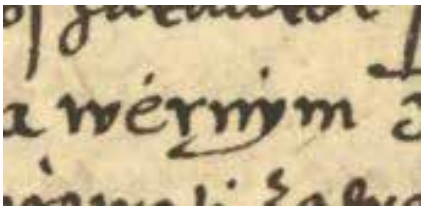
Obrázok 115 Pridávanie glyfov zo sady Unicode

Diakritika a ligatúry
Existujú dve možnosti pre spracovanie správneho prepisu znakov:

Možnosť 1 **Mierna normalizácia** podľa slovníka.

Hlavné pravidlo, ktoré sa tu uplatňuje: ak jasne vidíte základný znak glyfu (grafémy) a ak sa základný znak zároveň používa v slovníku na vyjadrenie tohto glyfu, zachovajte základný znak.

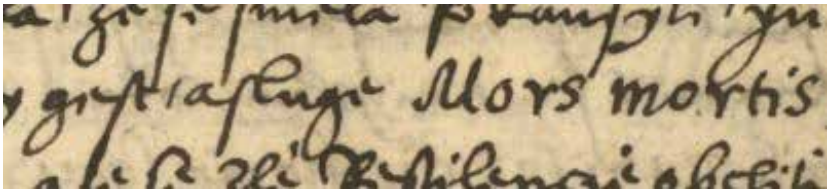
Príklad 1 litery *e* a *y* v latinskej minuskule sa v mnohých dokumentoch objavujú s diakritickými znamienkami, odlišnými od súčasného úzu (bodka, dvojbodka).



Obrázok 116 Slovo wernym zapísané novogotickou kurzívou. Litery sú prepísané bez diakritiky.

V jednoduchom prepise ich môžete prepísať ako latinskú minuskulnú literu *e* a *y*, keďže základný znak je stále jasne viditeľný.

Príklad 2 Latinská minuskulná litera *s* sa vo väčšine európskych historických písiem vyjadruje dvoma grafémami. Nachádzame preto jasný rozdiel medzi okrúhlym *s* a kurzívnym dlhým *ſ* (v štandardoch Unicode znak U+017F).

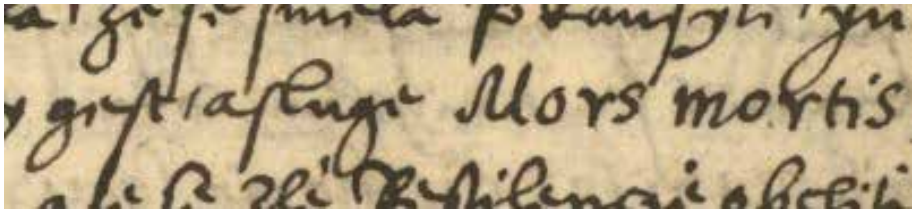


Obrázok 117 Jednoduchý prepis dlhého a okrúhleho s v slovách gest, asluge Mors mortis

Aj keď existuje jasný rozdiel, jednoduchý prepis by použil okrúhle *s* v obidvoch prípadoch.

Možnosť 2 **Paleografický prepis (transliterácia)**

Ortograficky vernému prepisu zodpovedá označenie transliterácia. Na platforme Transkribus sa pre všetky druhy prepisu konvenčne používa pojem transkripcia.



Obrázok 118 Paleografický prepis dlhého a okrúhleho s v slovách geſt, aſluge Mors mortis

V **tlačených textoch** (ktoré je tiež možné prepisovať) môže zohrať rolu prepisovanie ligatúr. Znovu možno použiť rovnaké pravidlo: Hoci sa špecifické kombinácie písmen napríklad *ft* alebo *ſt*, keď sa spájajú dve grafémy, dajú vyjadriť aj špecifickými znakmi Unicode, odporúčame ich prepisovať bez ligatúr podľa slovníka.

Interpunkčné znamienka

Interpunkčné znamienka sa prepisujú rovnakým spôsobom ako ostatné znaky. Použite príslušný znak na klávesnici. Na rozdiel od transkripčných pravidiel, ktoré interpunkčné znamienka pridávajú alebo vynechávajú podľa dnešného ponímania, odporúčame v prepise zachovať pôvodné znamienka. Napríklad dvojbodky sa v historických textoch často používajú na značenie skracovania slov. Mali by sa prepisovať ako dvojbodky.

Prepis novovekých dokumentov by mal odrážať predlohu, aj keď sa interpunkčné znamienko použilo spôsobom, ktorý nezodpovedá súčasnému úzu.

Prepis stredovekých dokumentov by nemal používať modernú interpunkciu. Vhodnejšie je vynechať všetky interpunkčné znamienka alebo použiť špecifické symboly zo sady Unicode.

Zásady prepisu v Transkribe

Je dôležité, aby bol jednoduchý prepis určený na trénovanie modelu PyLaia HTR konzistentný, vypracovaný jednotnou metodikou. Prepis má zodpovedať rukopisnej predlohe, vrátane pisárskych chýb.

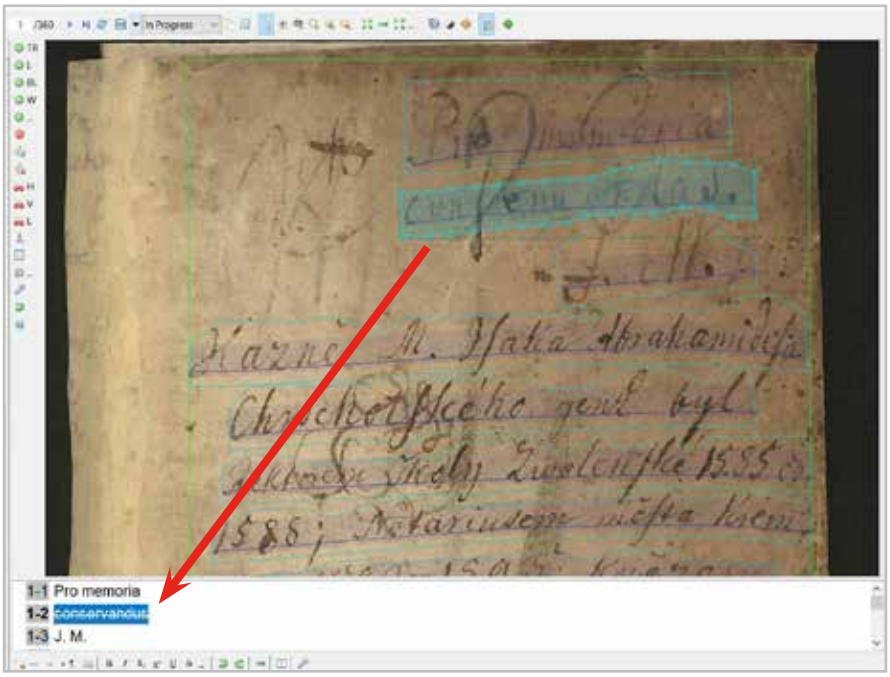
Slová by sa mali oddeľovať alebo spájať podľa predlohy.

Podľa rukopisu rozlišujte minuskulu a majuskulu. Ak literu nie je možné jasne rozlíšiť, rozhodnutie závisí od vás.

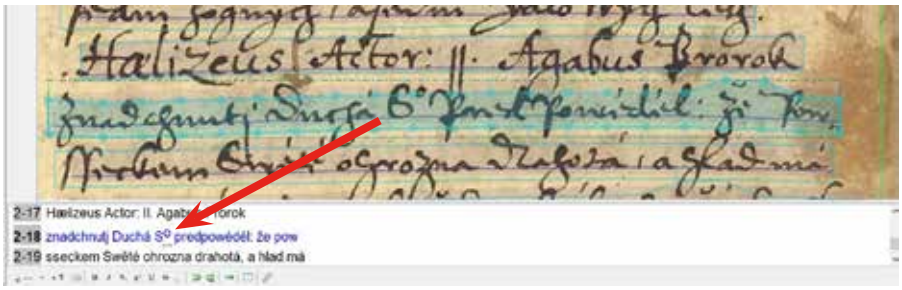
Slová s rozdeľovníkom na konci alebo uprostred riadku majú byť prepísané a rozdelené podľa pôvodného textu.

Prečiarknuté pasáže v texte môžete označiť tlačidlom Označiť ako prečiarknuté (*Tag as strikethrough*) na lište v poli textového editora. Podobne podčiarknuté pasáže môžete označiť tlačidlom Označiť ako prečiarknuté (*Tag as underlined*) na rovnakej lište.

Nadpísané pasáže textu (napr. značky skrátenia alebo interpunkciu) môžete označiť horným indexom na lište v poli textového editora tlačidlom *Tag as superscript*. Možnosti označenia dolného indexu ponúka tlačidlo *Tag as subscript*.



Obrázok 119 Označenie prepisovaného textu prečiarknutím



Obrázok 120 Označenie prepisovaného textu horným indexom

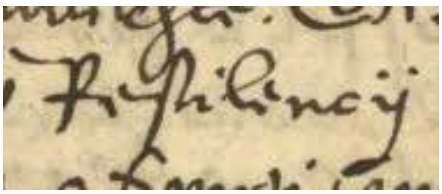
Odlišné typy a druhy písma (napr. gotické a humanistické) nie sú osobitne značené.

Zásady používania špeciálnych znakov

Skratky prepisujte podľa predlohy – nerozpisujte ich. Platí to pre historické spôsoby skracovania slov aj pre skratky používané v súčasnosti.

Diakritické znaky môžete vynechať (v prípade jednoduchého prepisu) alebo ich použiť podľa predlohy (v prípade transliterácie).

Častým prípadom je zamieňanie hlások *i* a *j*, ktoré je v rukopisoch náročné rozlíšiť najmä v majuskule. Zdvojenie znakov *ii* alebo *ij* sa prejavilo v používaní grafémy *ÿ*. Vkladajte ju pomocou virtuálnej klávesnice. Platí pritom odporúčanie, aby sa každý znak pre dostatočné osvojenie strojového učenia vyskytol v prepísanej vzorke aspoň 50-krát.



Obrázok 121 Dvojhláska v podobe samostatnej grafémy v zápise slova pestilencÿ

Ligatúry môžete rozpisovať, nie je potrebné používať pritom osobitné znaky ako pri skratkách. Ak sa rozhodnete ponechať ligatúru (napr. *æ*), mala by sa v prepísanej vzorke vyskytnúť v odporúčanom počte.

K formám zápisu hlásky *s*, okrem už spomínaného okrúhleho a dlhého *s*, patrí aj dvojité *s*, často v podobe ligatúry *ß*. Ostré *s* môžete prepisovať ako *ss* alebo použiť znak *ß*, ak sa v prepise vyskytuje 50-krát.

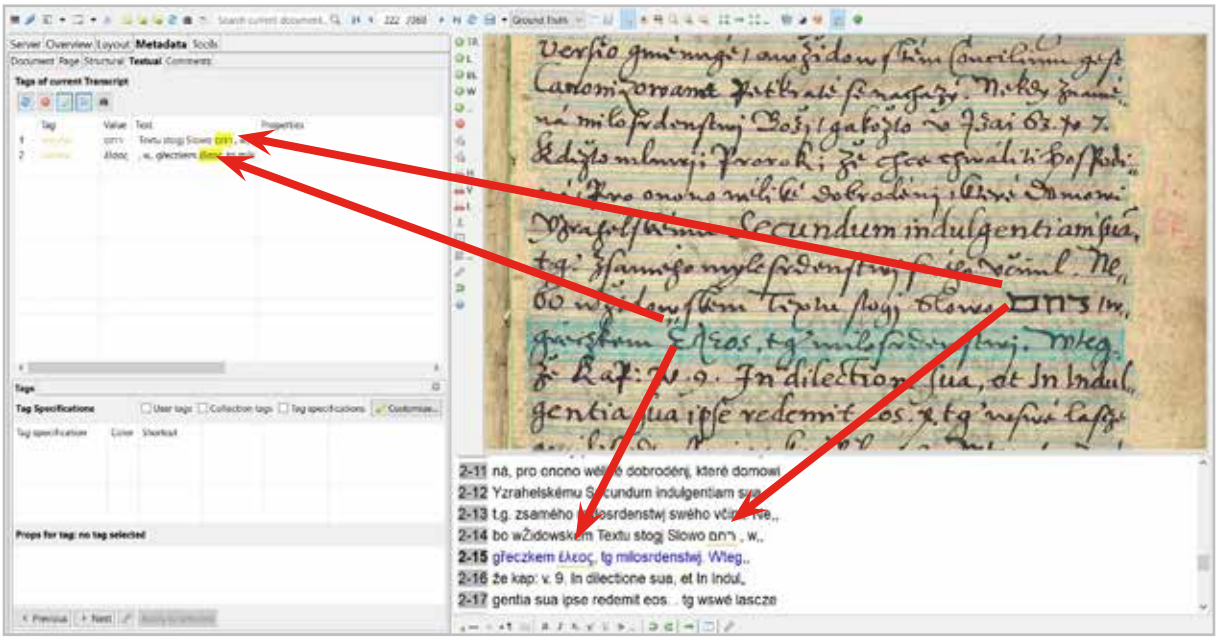
Tagovanie skratiek

Ak sa chcete venovať skratkám aj po prepise textu alebo ich chcete vynechať z trénovania modelu, môžete ich označiť tagom. Skrátené slovo alebo jeho časť s výskytom skratky označte v poli textového editora kliknutím na pravé tlačidlo myši. V ponuke sa zobrazí voľba Všetky tagy (*All tags*). Vyberte funkciu Skratka (*Abbrev*). Skrátené slovo zostane podčiarknuté červenou farbou. Označenie s hodnotou skratky sa zobrazí na paneli vľavo na záložke *Metadata*, časť *Textual*.

Nečitateľné miesta

V prípade nečitateľnej pasáže alebo pasáže zapísanej iným druhom písma ju môžete označiť pravým tlačidlom myši. Z voľby Všetky tagy (*All tags*) vyberte funkciu Nejasný (*Unclear*).

Pasáž bude podčiarknutá žltou farbou. Označenie sa opäť zobrazí v časti *Textual*. Takéto pasáže nemusia byť priamo zahrnuté do modelu (viac v kapitolách *5.2 Trénovanie modelu PyLaia* a *7.1.2 Ostatné textové tagy*).

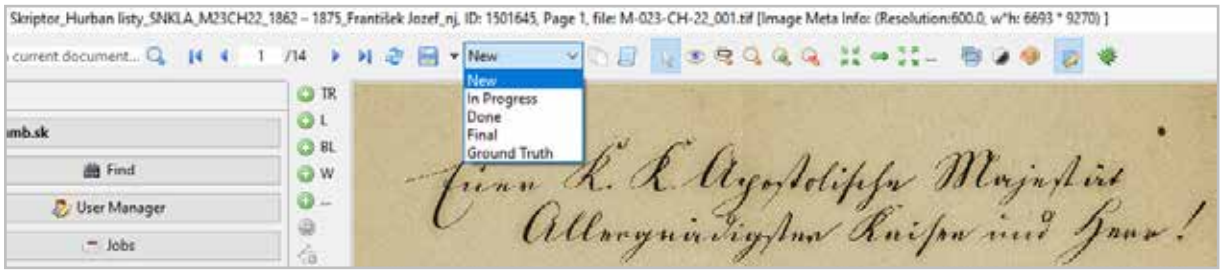


Obrázok 122 Označenie pasáží zapísaných hebrejčinou a gréčtinou funkciou Unclear

Stavy dokumentu a verzia *Ground Truth*

Editované alebo dokončené strany označte zodpovedajúcim atribútom v riadku na hornej lište v paneli nástrojov v hlavnom zobrazení. Môžete označiť tieto stavy strán dokumentu:

- nový (*New*) je automatické označenie strany po bezprostrednom po nahratí do expert klienta,
- prebiehajúci (*In Progress*) je označenie pre stranu, ktorú je stále možné prepisovať,
- hotový (*Done*) sa používa na označenie strany, ktorá je už prepísaná, ale ešte potrebuje kontrolu,
- finálna verzia (*Final*) označuje prepísanú a skontrolovanú stranu,
- „základná pravda“ (*Ground Truth*) takto označená strana je výsledná verzia prepisu, nemala by sa už meniť. Umožňuje pristúpiť k tvorbe modelu.



Obrázok 122a Označovanie stavu transkribovanej strany na paneli nástrojov

5. 2 Trénovanie modelu PyLaia

Pred spustením trénovania modelu je potrebné pripraviť si vzorku *Ground Truth* (viac v kapitole *5.1 Prepis dokumentu*), t. j. k originálu čo najpresnejší prepis (manuálny alebo automatic-

ko-manuálny), ktorý sa umelá inteligencia naučí „čítať“. V závislosti od typu prepisovaného dokumentu (tlač, rukopis) a počtu rúk (resp. meniaceho sa štýlu písania autora) sa odporúča trénovať model na 5 000 až 15 000 slovách, čo zodpovedá prepisu približne 25 až 75 strán:

- v prípade tlačeného textu na približne 5 000 slovách,
- v prípade rukopisného textu na aspoň 10 000 slovách pre každú ruku.

Ak chcete trénovať model na rozpoznanie troch rôznych „rúk“, mali by ste prepísať aspoň 30 000 slov, 10 000 slov pre každú ruku. Platí to aj v prípade jedného autora, ak sa jeho rukopis v priebehu života menil. Veľký model trénovaný na viac ako 100 000 slovách, ktorý obsahuje rôzne ruky z rovnakého obdobia a regiónu, by mal byť schopný rozpoznať aj rukopis, ktorý sa do tréningu nedostal (aj keď výsledky jeho prepisu môžu byť v porovnaní s trénovanými stranami o niečo horšie).

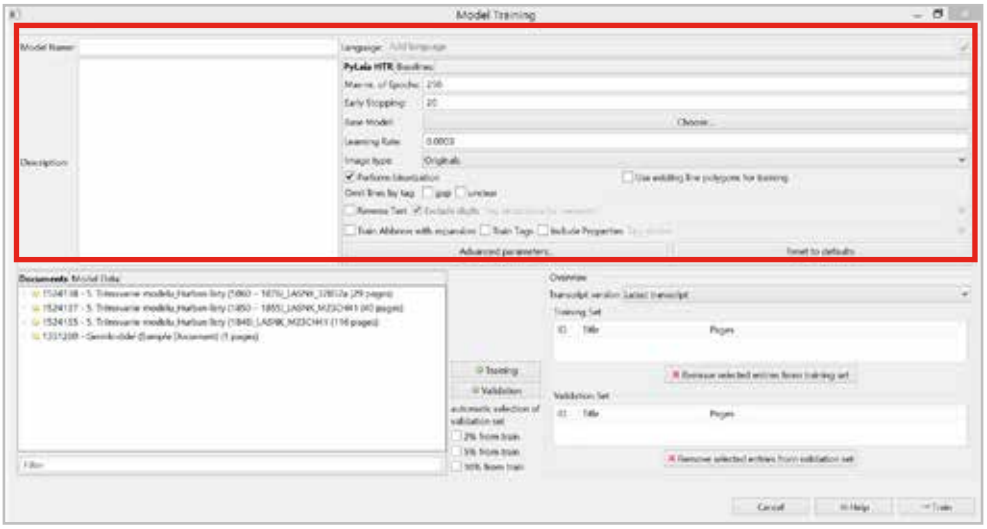
Je dôležité, aby strany vo vzorke *Ground Truth* boli **reprezentatívne**, t. j. aby obsahovali varianty všetkých typov písiem (resp. aj jazykov, abecied, no aj štýlov písania), ktoré má váš model byť schopný rozpoznať (čiže prepísať) súčasne. Strany zahrnuté do vzorky *Ground Truth* majú vplyv na kvalitu modelu.

5.2.1 Nastavenie parametrov pri trénovaní modelu PyLaia

Po príprave vzorky *Ground Truth* nasleduje spustenie **trénovania (nového) modelu**. Funkciu Trénovať nový model (*Train a new model*) nájdete na záložke Nástroje (*Tools*) v časti Trénovanie modelu (*Model Training*):

Obrázok 123 Záložka s funkciou trénovania (nového) modelu

Po jej výbere sa vám otvorí okno na trénovanie modelu. V hornej časti si nastavte vstupné údaje ako aj ďalšie parametre nástroja PyLaia, ktorými môžete zvýšiť funkčnosť a efektívnosť trénovaného modelu.



Obrázok 124 Okno na trénovanie modelu

Ako prvé uveďte **povinné údaje**:

- názov modelu (*Model Name*)
- jazyk dokumentu (*Language*): Jazyk(y) pridáte tak, že najprv kliknete na ikonku  , vpíšete názov jazyka do príslušného riadku (*Add Language*), potvrdíte dvojklikom, pomocou zeleného tlačidla plus pridáte do zoznamu (*Current languages*) a opätovne potvrdíte tlačidlom OK. V závislosti od jazyka vašej vzorky *Ground Truth* máte možnosť pridať jeden a viac jazykov.
- popis dokumentu (*Description*).

Následne podľa typu dokumentu a skúseností, aké nadobudnete pri práci s nástrojom PyLaia, môžete vyplniť ďalšie parametre (a podľa potreby meniť predvolené nastavenia):

a) štandardné parametre (PyLaia HTR)

- **maximálny počet cyklov** (*Max-nr. of Epochs*) predstavuje maximálny počet opakovaní tréningu, keď sa stroj „učí“ čítať cvičný súbor; t. j. pri každom cykle prečíta tú istú stranu a vyhodnotí ju. Na začiatok sa odporúča ponechať predvolené nastavenie (250 cyklov). Treba mať na pamäti, že zvyšovaním počtu cyklov sa aj proces trénovania predlžuje a naopak znižovaním zasa skracuje. Zvyšovanie počtu cyklov nemusí mať vplyv na výslednú úspešnosť modelu.
- **predčasné zastavenie** (*Early Stopping*) predstavuje minimálny počet opakovaní tréningu. Pre väčšinu modelov postačuje predvolené nastavenie (20 cyklov). Znamená to že, ak sa hodnoty modelu zlepšujú, trénovanie bude aj po dosiahnutí 20 cyklov pokračovať. Ak však už hodnoty nebudú vykazovať zlepšenie, trénovanie sa automaticky zastaví a vyhodnotí.
- **základný model** (*Base Model*) Ak chcete zefektívniť učenie, ako základný model si môžete vybrať jestvujúci, verejne dostupný model za predpokladu, že má podobné vlastnosti ako váš cvičný súbor. Keď do vami trénovaného modelu pridáte dáta základného modelu, umožní vám to začať s menšou vzorkou a za istých podmienok aj zlepšiť vami vytrénovaný model (viac o zdokonaľovaní modelu v kapitole *5.3 Vyhodnotenie úspešnosti modelu a jeho zdokonaľovanie*). Tabuľka s prehľadom základných modelov sa otvorí stlačením tlačidla *Choose…* Základný model pridáte dvojitým kliknutím na príslušný model a potvrdením tlačidlom OK.
- **rýchlosť učenia** (*Learning Rate*) Predvolená hodnota 0,0003, ktorú odporúčame ponechať, definuje, ako rýchlo bude učenie pri prechode od jedného cyklu k druhému prebiehať.
- **typ obrázka** (*Image Type*) Ak predbežné spracovanie trvá príliš dlho, zrýchliť ho môžete tak, že zmeníte typ obrázka – z originálu (*Originals*) na komprimovaný (*Compressed*).
- **vykonať binarizáciu** (*Perform binarization*) Táto možnosť je predvolená. Označenie možnosti zrušte len v prípade, že máte strany s rovnakou farbou pozadia.
- **na trénovanie použiť existujúci polygónový ťah** (*Use existing line polygons for training*) Túto možnosť označte, ak chcete počas trénovania zohľadniť existujúci, nie predvolený polygónový ťah.
- **vynechať riadky označené tagom** (*Omit lines by tag*) Túto možnosť označte, ak chcete z procesu trénovania vynechať riadky obsahujúce slová označené tagmi Medzera (*Gap*) alebo Nejasný (*Unclear*). Vynecháte tak nielen označené slovo, ale aj celý riadok, keďže trénovanie prebieha na úrovni riadkov.

- **obrátiť text** (*Reverse Text*) Túto možnosť označte, ak je smer písania na obrázku opačný ako pri prepise (napr. originál bol napísaný sprava doľava a prepísaný text zľava doprava). V tomto prípade sa tiež môžete rozhodnúť, či chcete z obráteného textu vylúčiť číslice (*Exclude digits*) alebo text označený tagom (*Tag exceptions for reversion*).
- **trénovať skratky** (*Train Abbrevs with expansion*) Túto možnosť použite, ak chcete dosiahnuť lepšie výsledky pri rozpoznávaní skratiek.
- **trénovať tagy/zahrnúť vlastnosti** (*Train Tags/Include Properties*) Tieto možnosti použite, ak chcete trénovať textové tagy a ich vlastnosti, ktoré sú súčasťou vzorky *Ground Truth*. Pomocou zeleného tlačidla plus pridávajte tagy, ktoré sa majú trénovať. Vďaka tejto funkcii môže model tagy v procese rozpoznávania generovať automaticky.
- **obnoviť predvolené nastavenia** (*Reset to defaults*) Túto možnosť vyberte, ak sa rozhodnete vrátiť k predvoleným nastaveniam.

b) pokročilé parametre (*Advanced Parameters...*)

Kliknutím na tlačidlo *Advanced Parameters…* sa otvoria **pokročilé parametre** nástroja PyLaia, rozdelené v stĺpcoch do troch skupín: predbežné spracovanie (*Preprocessing*), model (*Model*) a trénovanie (*Training*). Ponechanie preddefinovaných nastavení odporúčame zvážiť v závislosti od špecifík vášho dokumentu. Ak napríklad trénujete tlačené dokumenty obsahujúce kurzívu, zrušte predvolené označenie funkcie *Deslant*, ktorá slúži na vyrovnávanie kurzívneho písma. Štyri parametre v stĺpci *Preprocessing* však nemeňte (*Moment normalization; Features parallelogram; Features surrounding polygon* a *Features surrounding polygon dilate*).

5.2.2 Spustenie trénovania modelu PyLaia

V spodnej časti okna na trénovanie modelu sa na záložke Dokumenty (*Documents*) nachádza pripravená vzorka *Ground Truth*. Príprava trénovania modelu spočíva nielen v nastavení príslušných parametrov, ale aj v rozdelení vzorky do dvoch súborov:

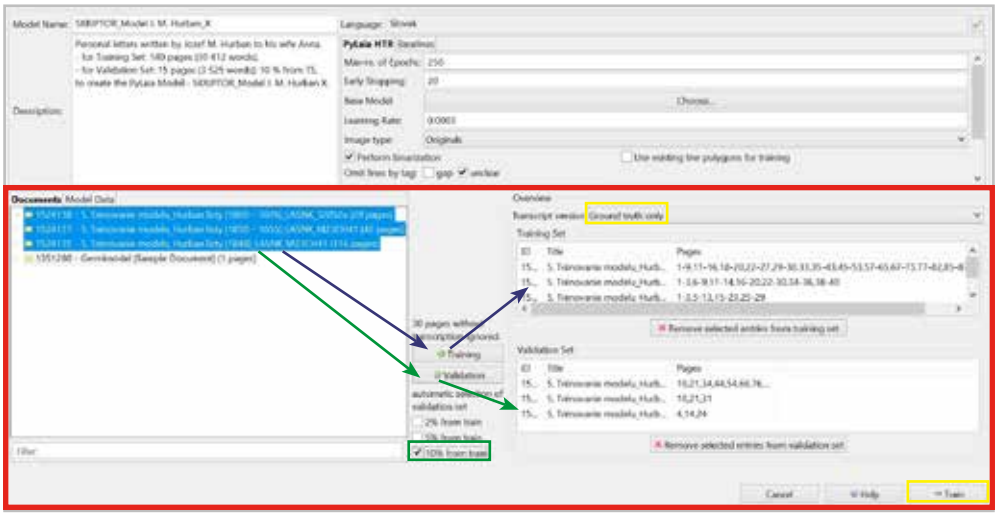
- Do **cvičného súboru** (*Training Set*) vyberáte strany, na ktorých sa model môže vytrénovať. Na cvičnom súbore sa stroj „učí“, pri každom cykle „prečíta“ rovnakú stranu, pričom chybne prečítané znaky pri každom nasledujúcom cykle vyradí.
- Do **overovacieho súboru** (*Validation Set*) vyberáte strany, na ktorých sa presnosť vytrénovaného modelu automaticky overí (odskúša). V porovnaní s cvičným súborom je preto menší, spravidla 10 % z celkovej vzorky *Ground Truth*. Na druhej strane overovací súbor by mal byť reprezentatívny, t. j. mal by obsiahnuť príklady všetkých písmen, jazykov a iných atribútov zahrnutých v cvičnom súbore. V opačnom prípade, čiže ak je overovací súbor príliš homogénny, výkon modelu môže byť nízky, prípadne skreslený.

Označené súbory alebo samostatné strany (po rozbalení priečinka), ktoré chcete pridať do cvičného súboru alebo do overovacieho súboru, vyberajte pomocou zelených tlačidiel na trénovanie (+*Training*) a overovanie (+*Validation*).

Ak sa rozhodnete pre automatický výber overovacieho súboru (*Automatic selection of validation set*), označte najskôr stránky, ktoré chcete pridať do cvičného súboru, potom označte príslušné percento strán (2 %, 5 % alebo 10 %), ktoré chcete priradiť do overovacieho súboru, a potom stlačte tlačidlo +*Training*.

Ak chcete niektoré strany z cvičného alebo overovacieho súboru odobrať, príslušnú stranu označte, a potom zvoľte ×*Remove selected entries from training/validation set*.

Pri presune strán do oboch súborov sa odporúča vybrať si verziu prepisu (*Transcript version*) – *Ground truth only*, a to ako poistku, že sa do nich nedostanú strany s iným príznakom (napr. *In Progress*).



Obrázok 125 Ukážka nastavenia parametrov a rozdelenia vzorky Ground Truth pri trénovaní modelu

Trénovanie modelu PyLaia spustíte tlačidlom Trénovať (*Train*). Otvorí sa okno s prehľadom dát o cvičnom a overovacom súbore (*Dataset Overview*). Ak sú pre vás tieto dáta postačujúce (napr. počet slov v cvičnom alebo overovacom súbore), stlačte tlačidlo Spustiť trénovanie (*Start training*). Proces spracovania údajov a priebeh trénovania si môžete skontrolovať v zobrazení hlavného menu pod ikonkou (*Show jobs*).

Pri tvorbe modelu PyLaia sa zo vzorky *Ground Truth* zároveň generujú tzv. **jazykové modely**, ktoré sa môžu použiť pri transkripcii textov. Pomáhajú určovať pravdepodobnosť poradia slov alebo frekvenciu ich výskytu a kombinácie v istom kontexte. Ich použitie môže mať vplyv na zlepšenie výsledkov transkripcie.

5.3 Vyhodnotenie úspešnosti modelu a jeho zdokonaľovanie

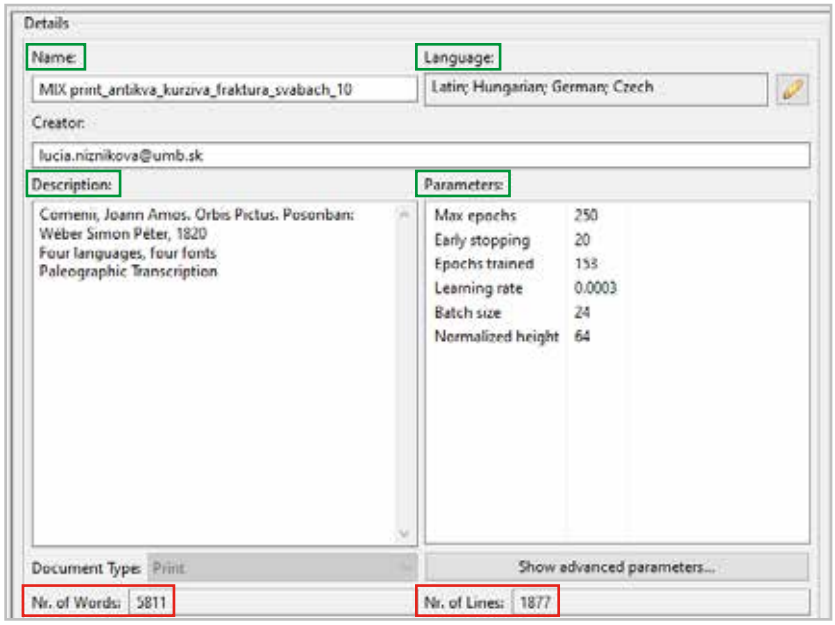
5.3.1 Hodnotenie úspešnosti modelu

Po vytrénovaní modelu expert klient ponúkne výsledok v podobe grafu a percentuálneho vyjadrenia chybovosti znakov v automaticky prepísanom texte. Výsledok je dostupný na záložke Nástroje (*Tools*) v časti Zobraziť modely (*View Models*). V ľavej časti okna sa zobrazí zoznam dostupných modelov vrátane vášho modelu. Po kliknutí na príslušný model sa na pravej strane zobrazí vyhodnotenie, ktoré má dve časti.

Prvá časť vyhodnotenia obsahuje vstupné údaje o transkribovanom dokumente, ktoré zadávate pred spustením modelu:

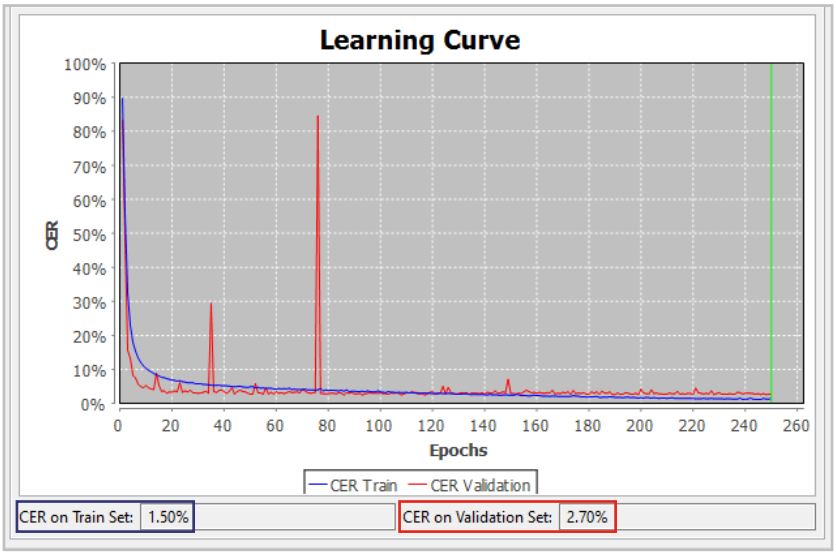
- názov modelu (*Name*),
- jazyk dokumentu (*Language*),
- popis dokumentu (*Description*),
- súhrn nastavení pred spustením trénovania (*Parameters*).

Softvér prepočíta aj počet slov (*No. of Words*) a počet riadkov (*No. of Lines*), ktoré vstupovali do trénovania. Je to dobrá pomôcka, keďže na základe týchto údajov viete posúdiť, či bol rozsah cvičného a overovacieho súboru dostatočný.



Obrázok 126 Vstupné údaje o transkribovanom dokumente

Druhá časť vyhodnotenia obsahuje grafické zobrazenie procesu trénovania, t. j. krivku učenia (*Learning Curve*) a chybovosť znakov v cvičnom a overovacom súbore na úrovni dokumentu ako celku. Graf zobrazuje presnosť vášho modelu.



Obrázok 127 Grafické zobrazenie procesu trénovania

Os y predstavuje **mieru chybovosti znakov CER** (*Character Error Rate*) a udáva sa 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**

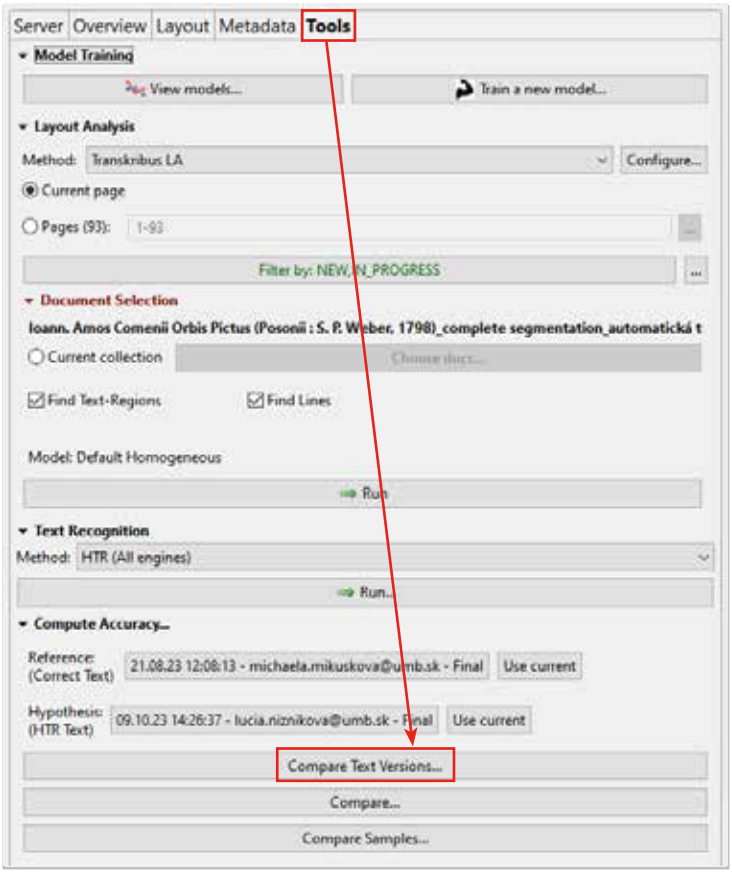
Každá malá chyba pri prepise je štatisticky plnohodnotnou chybou. To znamená, že chýbajúca čiarka, *u* namiesto *v*, *á* namiesto *ä*, medzera navyše alebo veľké písmeno namiesto malého sa počítajú ako chyby.

Os x predstavuje cykly, t. j. priebeh tréningu. Počas procesu trénovania vykonáva Transkribus vyhodnotenie po každom cykle. Model na obrázku 126 bol vytrénovaný pri počte 101 cyklov. V tomto prípade bol maximálny počet cyklov nastavený na 220, ale trénovanie sa automaticky zastavilo pri 101, pretože model sa už nezlepšoval.

V grafe sú zobrazené dve čiary, jedna modrá a druhá červená. Modrá čiara predstavuje priebeh trénovania (učenia). Červená čiara predstavuje priebeh vyhodnocovania na overovacom súbore.

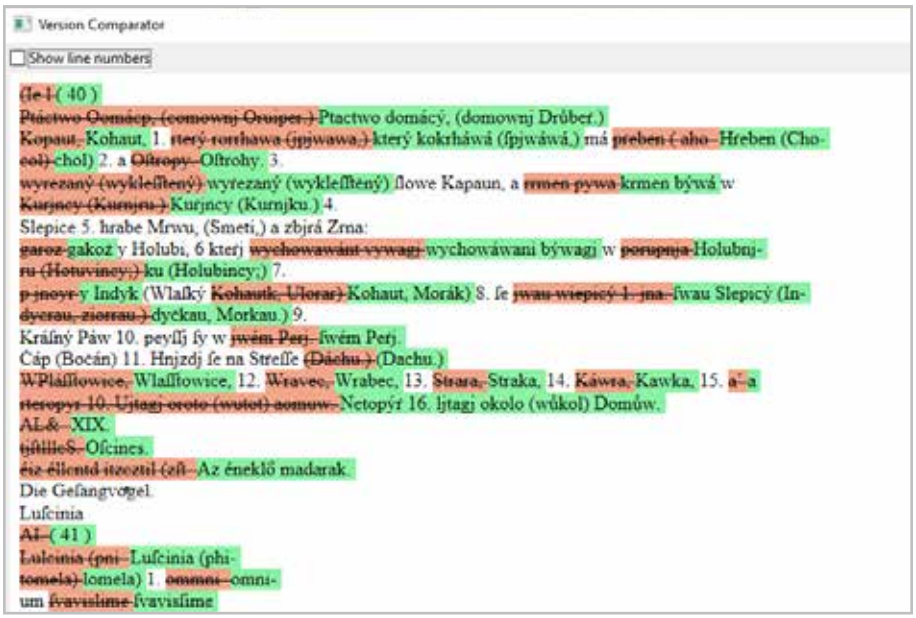
V tejto fáze ponúka Transkribus **dve hodnoty miery chybovosti znakov** – na cvičnom súbore (*CER on Train Set*) a na overovacom súbore (*CER on Validation Set*). Miera chybovosti znakov na overovacom súbore je z hľadiska hodnotenia úspešnosti modelu dôležitejšia, pretože ukazuje, ako si model poradil so stranami, na ktorých nebol vycvičený. Hodnoty *CER on Validation Set* 5 % a menej možno považovať za vynikajúci výsledok automatického prepisu, hodnoty do 10 % za uspokojivé.

Odlišne sa vyhodnocuje úspešnosť modelu na úrovni jednotlivých strán. Prvý spôsob je porovnanie textových verzií. Na záložke Nástroje (*Tools*) vyberte funkciu Porovnať textové verzie (*Compare Text Versions…*) Získate tak podrobný prehľad toho, čo model prepísal správne a kde v porovnaní s verziou *Ground Truth* urobil chybu.



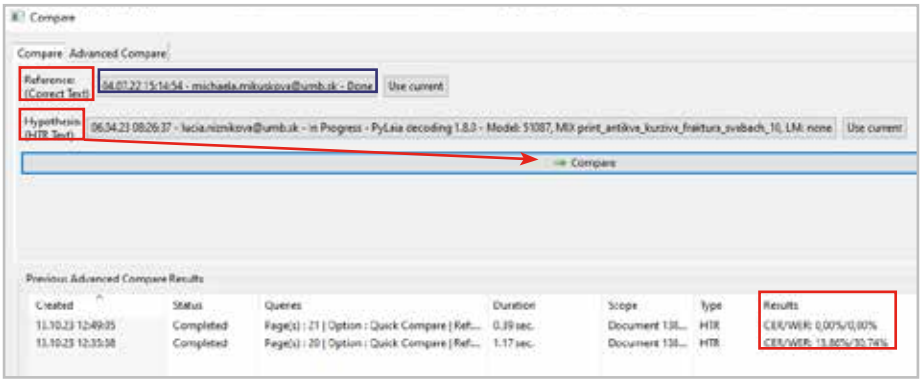
Obrázok 128 Porovnanie textových verzií

Je dôležité uvedomiť si, že ak je chybný čo len jeden znak, červenou farbou sa označí celé slovo. Slovo zvýraznené zelenou farbou je zobrazené tak, ako je prepísané vo verzii *Ground Truth*. Neoznačené slová sú tie, ktoré model rozpoznal totožne s *Ground Truth*.



Obrázok 129 Ukážka chybovosti v trénovanom modeli

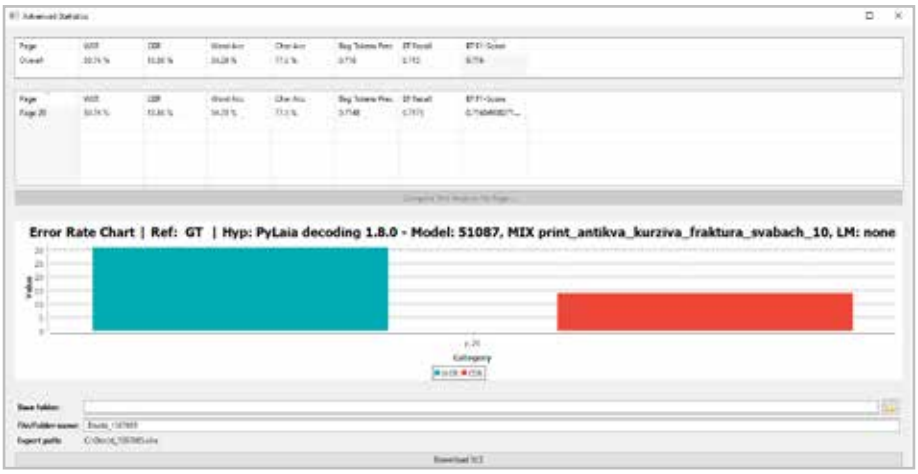
Najrýchlejší spôsob, ako si štatisticky overíte **chybovosť na úrovni strán**, je použitie funkcie Porovnať (*Compare…*) na záložke Nástroje (*Tools*). Na začiatku je dôležité uistiť sa, že ste v hornej časti okna vybrali správne verzie dokumentu, ktoré chcete porovnať – manuálne prepísaný text, resp. *Ground Truth* (*Reference, Correct Text*) a text prepísaný automaticky (*Hypothesis, HTR Text*). Potom stlačte tlačidlo Porovnať (*Compare*). Výsledok sa po niekoľkých sekundách zobrazí v dolnej pravej časti okna.



Obrázok 130 Výsledok chybovosti na úrovni strán

V stĺpci Výsledky (*Results*) sa pri každej porovnanej strane zobrazí nielen miera chybovosti znakov CER, ale aj **miera chybovosti slov WER** (*Word Error Rate*)**.** Na obrázku je miera chybovosti znakov strany *Page1* 1,26 %, čo znamená, že 98,74 % znakov v automatickom prepise je správnych. Podobne miera chybovosti slov 4,56 % znamená, že 95,44 % slov na strane je prepísaných správne. Musíme si však uvedomiť, že najviac chybných slov obsahuje väčšinou chyby späté s interpunkciou (chýbajúca alebo nadbytočná bodka, čiarka, dvojbodka a pod.), resp. s diakritikou (krátka samohláska namiesto dlhej, resp. naopak), ktoré nemajú takmer žiadny vplyv na zrozumiteľnosť textu. To je dôvod, prečo sa preferuje sledovanie chybovosti znakov CER pred chybovosťou slov WER.

Dvojitým kliknutím na dátum a čas v stĺpci *Created* (vľavo) sa automaticky otvorí okno rozšírených štatistík (*Advanced Statistics*). Tu získate podrobnejšie údaje a hodnoty, a výsledky môžete exportovať do súboru Excel (*Download XLS*).



Obrázok 131 Okno rozšírených štatistík

5.3.2 Zdokonaľovanie modelu

Ak výsledné hodnoty chybovosti modelu nevyšli podľa vašich predstáv a máte pocit, že model sa dá vylepšiť, môžete využiť niektorý z týchto postupov.

Postup 1 **Oprava manuálnej transkripcie**

Pri podrobnej analýze automatickej transkripcie po vytrénovaní modelu možno zistíte, že softvér upozornil aj na vaše vlastné chyby pri manuálnej transkripcii. V tejto fáze môžete manuálny prepis strán opätovne skontrolovať a opraviť prípadné nezrovnalosti. Vždy dbajte na to, aby strany vo vzorke *Ground Truth* boli prepísané presne a bezchybne, v opačnom prípade môže každý nesprávne prepísaný znak a akékoľvek odchýlenie od originálu negatívne ovplyvniť výsledok trénovania modelu.

Postup 2 **Zvýšenie počtu slov v cvičnom súbore**

V niektorých prípadoch odporúčaný minimálny počet slov zahrnutých do trénovania modelu (viac v kapitole *5.2 Trénovanie modelu PyLaia*) nie je dostačujúci, preto je potrebné rozsah zväčšiť. Ide o metódu priebežného zvyšovania strán v cvičnom súbore o automaticky transkribované strany (t. j. na základe aktuálnej verzie modelu). Automaticky prepísané strany sa následne opravia na úroveň *Ground Truth* a pripoja sa k predošlým, „bezchybne“ pripraveným stranám, aby sa mohlo spustiť trénovanie nového, väčšieho súboru. Keďže vo väčšine prípadov platí pravidlo, že čím vyšší je výskyt znaku v cvičnom súbore, tým lepšie sa ho stroj naučí rozpoznávať, týmto postupom môžete dosiahnuť zníženie miery chybovosti na úrovni znakov v overovacom súbore aj o niekoľko percentuálnych bodov.

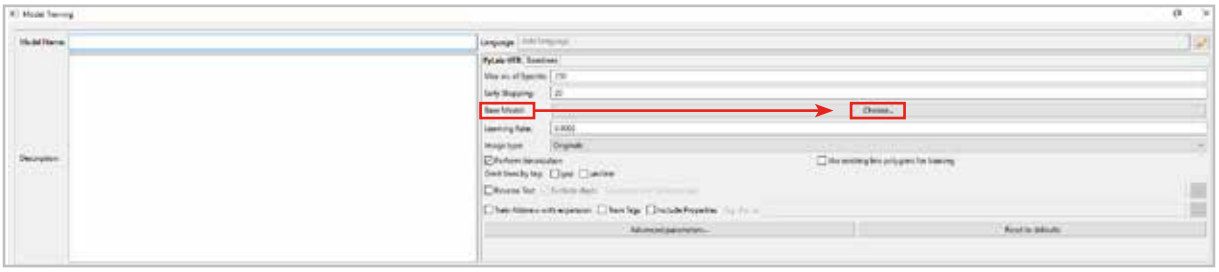
Postup 3 **Použitie základného modelu (Base Model)**

Z doterajšej praxe sa ako najefektívnejší spôsob zdokonaľovania modelu javí využitie tzv. základného modelu (*Base Model*). Základný model si pamätá, čo sa naučil. Preto každé nové trénovanie (teoreticky) zlepšuje jeho kvalitu. Nový model sa učí od svojho predchodcu, a tým sa stáva lepším. Trénovanie pomocou základného modelu je preto mimoriadne vhodné pre veľké

všeobecné modely, ktoré sa priebežne vyvíjajú počas dlhého obdobia. Pri tomto postupe máte na výber dve možnosti:

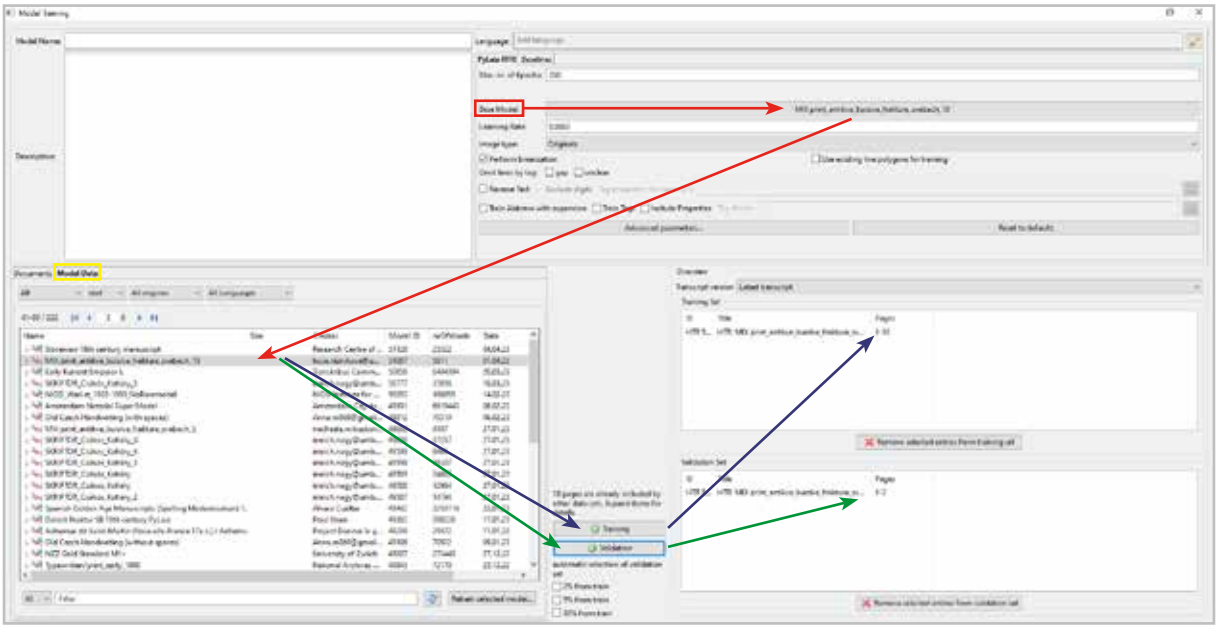
- použiť najlepšiu verziu vlastného modelu,
- použiť model iného používateľa, ktorý je verejne dostupný (najmä tento je vhodný aj pre malé modely trénované na dokumentoch s menším rozsahom strán – slúži ako kompenzácia menšieho počtu tréningových údajov spôsobených nedostatočným rozsahom prepisovaného dokumentu).

Ak chcete spustiť trénovanie pomocou základného modelu, na záložke Nástroje (*Tools*) vyberte funkciu Trénovať nový model (*Train* a new model) a okrem bežných nastavení zvoľte príslušný základný model.



Obrázok 132 Výber základného modelu (Base Model)

Potom na záložke *Model Data* vložte cvičný súbor a overovací súbor základného modelu, ako aj cvičný a overovací súbor novej vzorky *Ground Truth*. Spustite trénovanie.



Obrázok 133 Trénovanie pomocou základného modelu

Upozorňujeme že, výberu základného modelu od iného používateľa musí predchádzať dôkladná analýza postupov a metód, ktoré boli na jeho vytrénovanie použité. Ak takýto základný model nebol trénovaný na podobnom type rukopisného alebo tlačeného fontu a nemá špecifické znaky, ktoré obsahuje váš dokument, na automatickú transkripciu je nevhodný.

Na záver je dôležité podotknúť, že na výsledné hodnoty vytrénovaného modelu vplýva viacero faktorov:

a) Faktory, ktoré môžete ovplyvniť:
   - kvalita digitalizátu – preexponované alebo inak nekvalitné snímky nahradiť lepšími zábermi,
   - charakter textov, ktoré sa počas trénovania rozhodnete vložiť do overovacieho súboru – či už ide o mieru ich reprezentatívnosti, kvalitu alebo počet znakov na príslušnej strane (napr. v poslednom prípade platí, že čím menej znakov na strane, tým väčšie percento chybovosti),
   - kvalita manuálnej transkripcie textu – správnosť prepisu je základom správneho učenia sa stroja a každý nesprávne prepísaný znak znižuje kvalitu vytrénovaného modelu, resp. zvyšuje výsledné hodnoty miery chybovosti znakov.

b) Faktory, ktoré nedokážete ovplyvniť:
   - kvalita originálneho dokumentu – ak je originálna tlač nekvalitná, písmo nevýrazné (málo sýte), text obsahuje zásahy perom/ceruzkou (podčiarknuté riadky, škrtance, nadpísané slová a pod.), machule a iné nečistoty na papieri,
   - pri rukopisných textoch platí dvojnásobne, že odchýlky v rukopise (napr. zmena štýlu písania, častý výskyt autorských korektúr, hromadné uvádzanie číselných údajov) môžu negatívne ovplyvniť výslednú úspešnosť modelu.

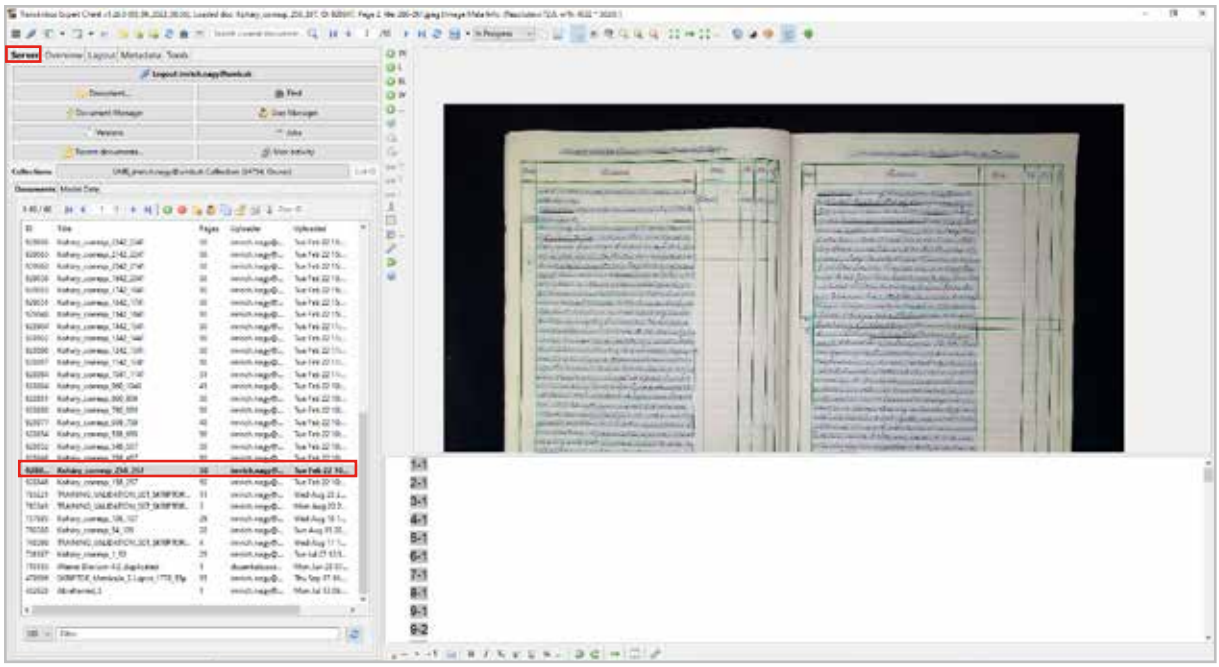
Z vyššie uvádzaných dôvodov je dôležité pri príprave vzorky *Ground Truth* zvoliť čo najtypickejšie, reprezentatívne a nepoškodené strany z rukopisu/tlače.

6 Priebeh automatickej transkripcie v expert klientovi

Automatická transkripcia dokumentu je zavŕšením práce v expert klientovi, od ktorého očakávame výstup v podobe jeho zrozumiteľného a všestranne použiteľného digitálneho prepisu. Pred samotnou realizáciou automatickej transkripcie ešte raz skontrolujte, či ste vykonali všetky prípravné kroky:

- dokument mám zdigitalizovaný,
- digitalizáty som importoval na platformu Transkribus expert klient,
- vykonal som segmentáciu textu,
- mám model pre automatickú transkripciu dokumentu (vytrénoval som vlastný model, resp. chcem použiť adekvátny model z portfólia voľne dostupných modelov v expert klientovi)

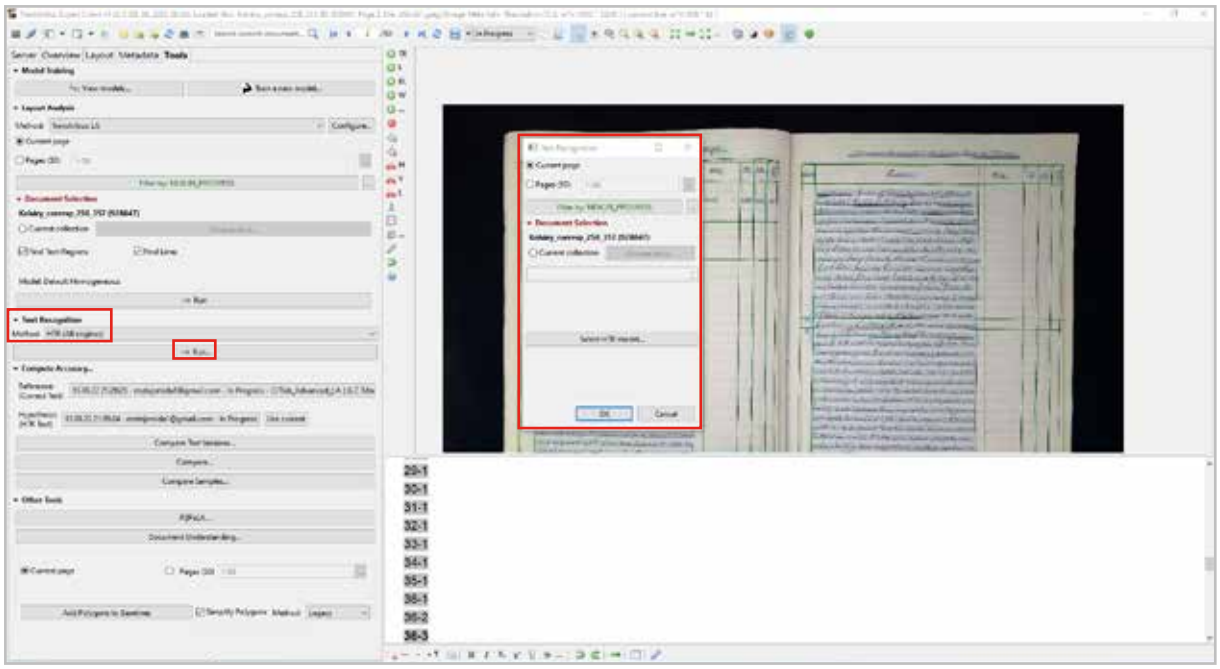
Výber dokumentu na automatickú transkripciu
Na záložke *Server* vyberte požadovaný dokument.



Obrázok 134 Výber dokumentu na automatickú transkripciu

Voľba nástroja automatickej transkripcie
Na hlavnej lište otvorte záložku Nástroje (*Tools*), na ktorej budete v tomto kroku používať voľbu Rozpoznávanie textu (*Text Recognition*).

V ponuke Metóda (*Method*) ponechajte voľbu *HTR (All engine)* na automatickú transkripciu rukopisného textu. Ak chcete transkribovať tlačený text (vhodné napríklad pre strojopisný dokument), zvoľte druhú možnosť Transkribus OCR *(Block-segmentation + Transkribus-Print-M1 model)*. Stlačte tlačidlo Spustiť (*Run*), ktorým otvoríte samostatné okno rozpoznávania textu (*Text Recognition*).



Obrázok 135 Voľba nástroja automatickej transkripcie a otvorenie samostatného okna pre rozpoznávanie textu (Text Recognition)

Výber snímok na automatickú transkripciu

V hornej časti okna si nastavte **výber snímok**, na ktorých chcete spustiť automatickú transkripciu. Máte niekoľko možností:

Ak ponecháte zakliknutú predvolenú **aktuálnu stranu** (*Current page*), automatická transkripcia sa spustí len na tej snímke dokumentu ktorú máte zobrazenú v hlavnom okne (v našom prípade je to snímka 2 z celkového počtu 50 snímok).

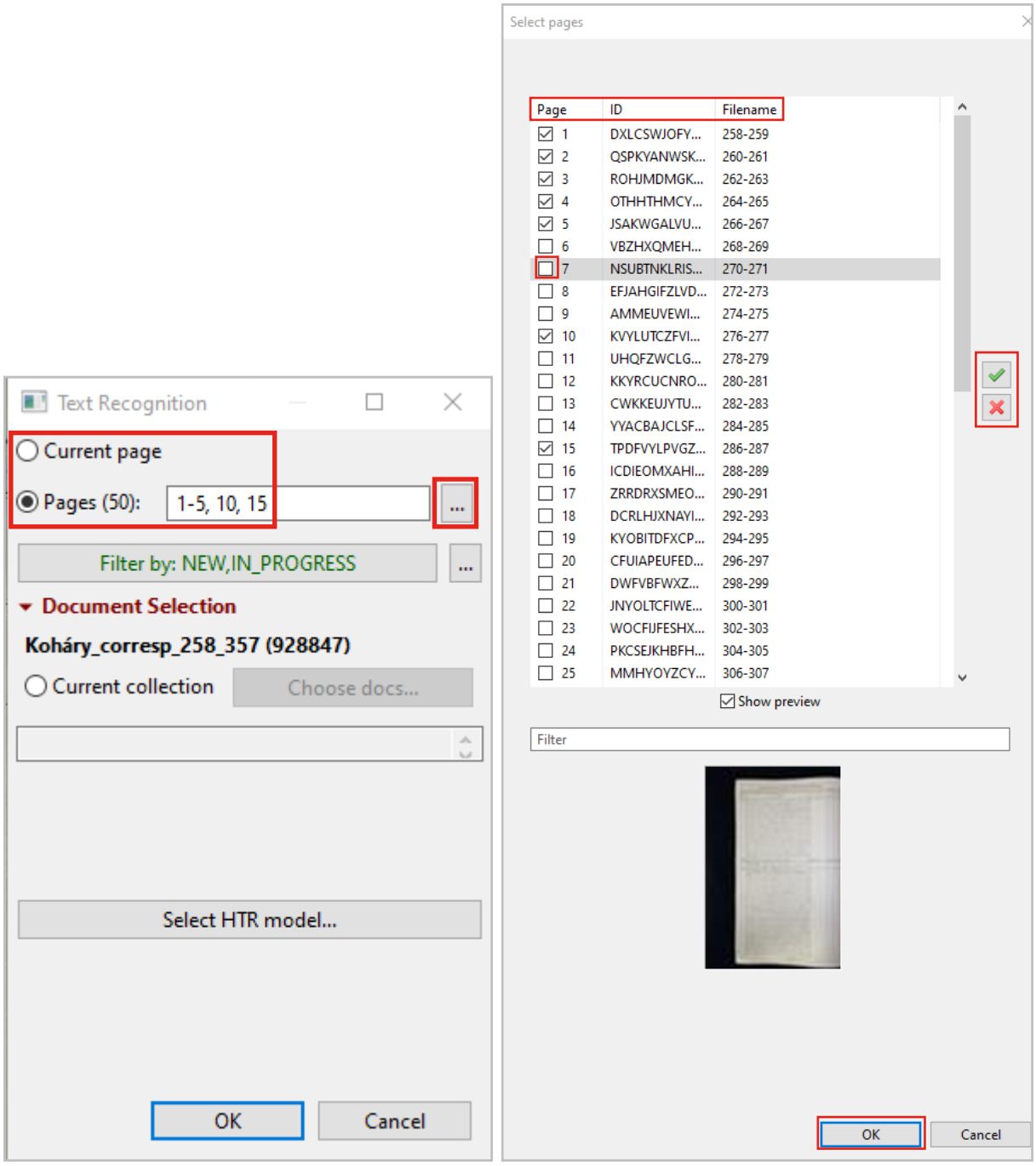
Poznámka: Pri digitalizovaní dokumentu sa na jeden záber spravidla snímajú dve strany, ako je to aj v našom prípade. Termín page, resp. pages teda označuje tieto snímky. Ak si vyberiete na automatickú transkripciu jednu snímku, dostanete vo výstupe prepis obidvoch strán dokumentu, ktoré sú zachytené na konkrétnej snímke.

Ak použijete voľbu **Strany** (*Pages*), otvorí sa okno s rozsahom snímok celého dokumentu (v našom prípade 1 – 50). Tento rozsah môžete podľa potreby prepísať, prípadne uviesť konkrétne snímky (napr. 1 – 5, 10, 15).

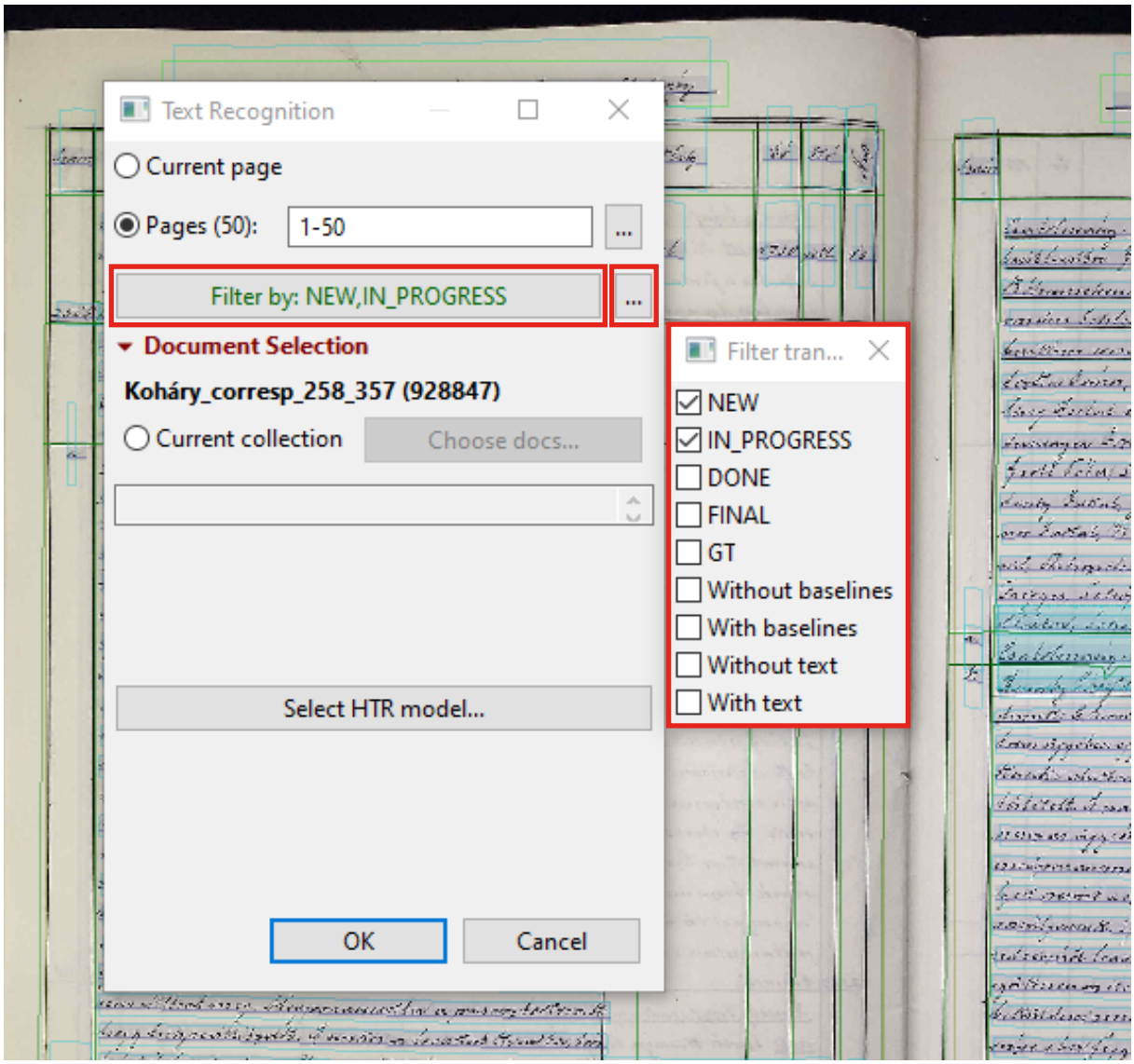
Ak kliknete na **tlačidlo s tromi bodkami** vedľa riadku s rozsahom snímok, otvorí sa nové dialógové okno, v ktorom môžete skontrolovať váš výber:

- jednotlivé snímky sú v ňom identifikované poradovým číslom (*Page*), identifikátorom prideleným Transkribom (*ID*) a pôvodným menom, ktorým bola snímka označená pred importom do Transkribu (*Filename*),
- v spodnej časti dialógového okna sa súčasne zobrazí náhľad časti snímky, ktorý ste označili kliknutím myšou,
- podľa potreby môžete výber konkrétnych snímok korigovať zakliknutím okienka pri poradovom čísle, resp. opätovným kliknutím svoj výber vymazať,
- na výber všetkých snímok môžete použiť tlačidlo so zeleným začiarknutím,
- na zrušenie všetkých vybratých snímok môžete použiť tlačidlo s červeným krížikom,
- úpravy potvrdíte tlačidlom OK v spodnej časti dialógového okna, resp. zrušíte tlačidlom *Cancel*.

Výber snímok si môžete uľahčiť aj ich filtrovaním pomocou tlačidla Filtrovať podľa (*Filter by:*) Filter nastavíte kliknutím na tlačidlo s tromi bodkami umiestnené vedľa tlačidla filtrovania a následným výberom nastaveného stavu (príznaku) snímky (*New*, *In Progress*, *Done*, *Final*, *Ground Truth*) alebo charakteristiky snímky (*With/Without Baselines*, *With/without text*).



Obrázok 136 Výber snímok s otvoreným dialógovým oknom, v ktorom je podrobný súpis jednotlivých snímok aj s čiastočným náhľadom na digitalizát



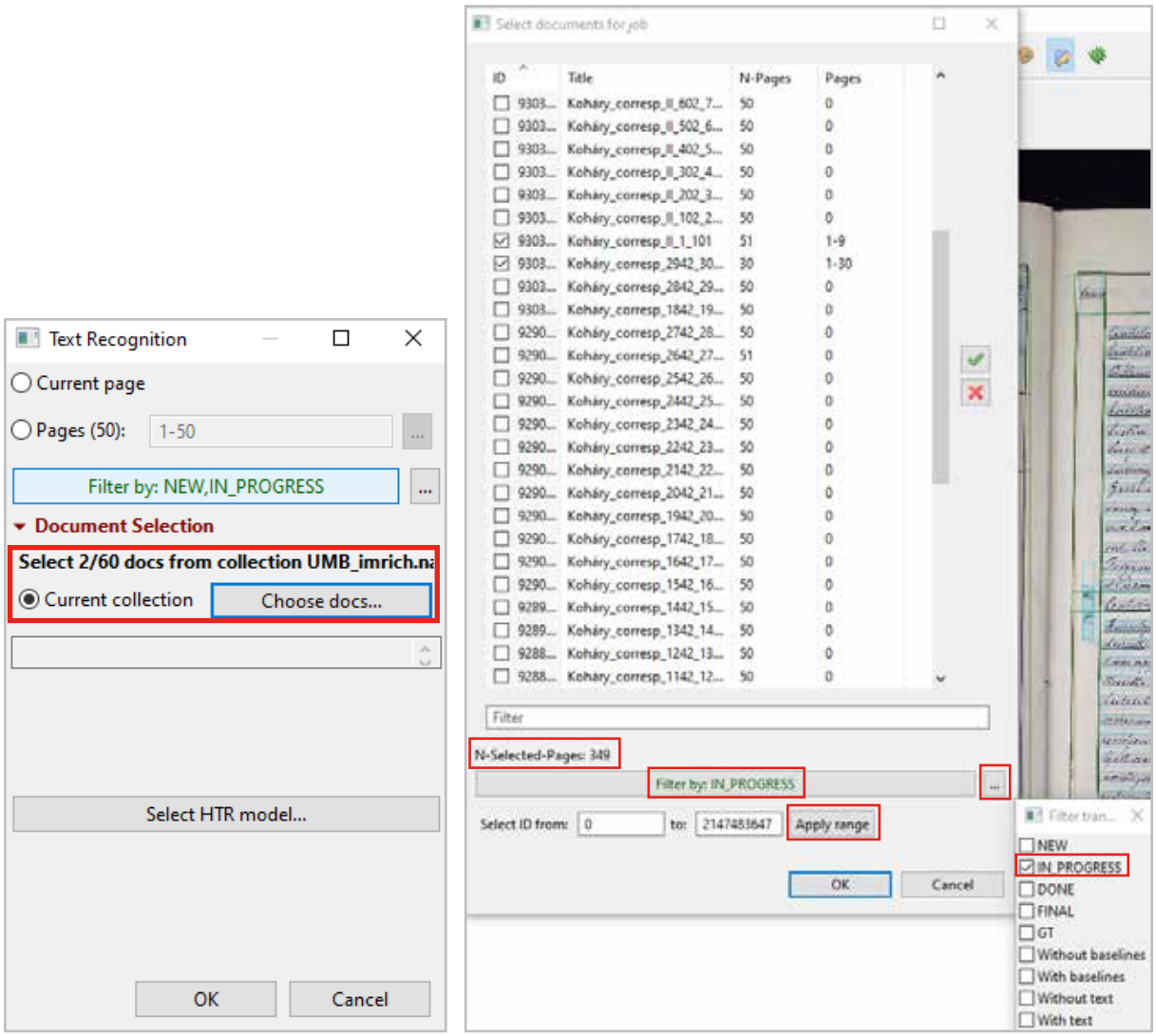
Obrázok 137 Nastavenie filtra na výber snímok

POZOR! Pri výbere rozsahu snímok na automatickú transkripciu je potrebné vziať do úvahy, že za prepis každej snímky sa odpočítajú kredity z konta používateľa. Preto ak ešte nemáte odskúšaný model automatickej transkripcie (napr. pri jeho prvej aplikácii na zvolený dokument), odporúčame vybrať iba obmedzený rozsah snímok (vzorku dokumentu).

Ak máte odskúšaný model automatickej transkripcie a pripravené celé súbory dokumentov, na ktoré ho chcete aplikovať, môžete využiť **ponuku na výber dokumentov** (*Document Selection*) vo vašej zbierke (*Current collection*):

- kliknutím na tlačidlo Vybrať dokumenty (*Choose docs*) sa otvorí nové dialógové okno, ktoré obsahuje zoznam všetkých dokumentov, ktoré ste importovali do svojej zbierky (konta),
- pri každom dokumente je uvedený identifikátor pridelený softvérom (ID), pôvodný názov dokumentu, počet a rozsah snímok,
- výber dokumentu urobíte zakliknutím políčka vedľa ID dokumentu, v spodnej časti sa zobrazuje automatický súčet snímok zo všetkých vybratých dokumentov,
- opäť máte v ponuke aj filter s možnosťou nastavenia stavu (príznaku), resp. charakteristiky snímky cez tlačidlo s tromi bodkami,

- po aplikácií takto nastaveného filtra kliknutím na tlačidlo *Apply range* a následne na tlačidlo *Filter by:* sa automaticky zvolia všetky snímky zo všetkých dokumentov vo vašej zbierke, ktoré spĺňajú nastavenia filtra,
- výber potvrďte tlačidlom OK v spodnej časti dialógového okna, resp. zrušte tlačidlom *Cancel*.

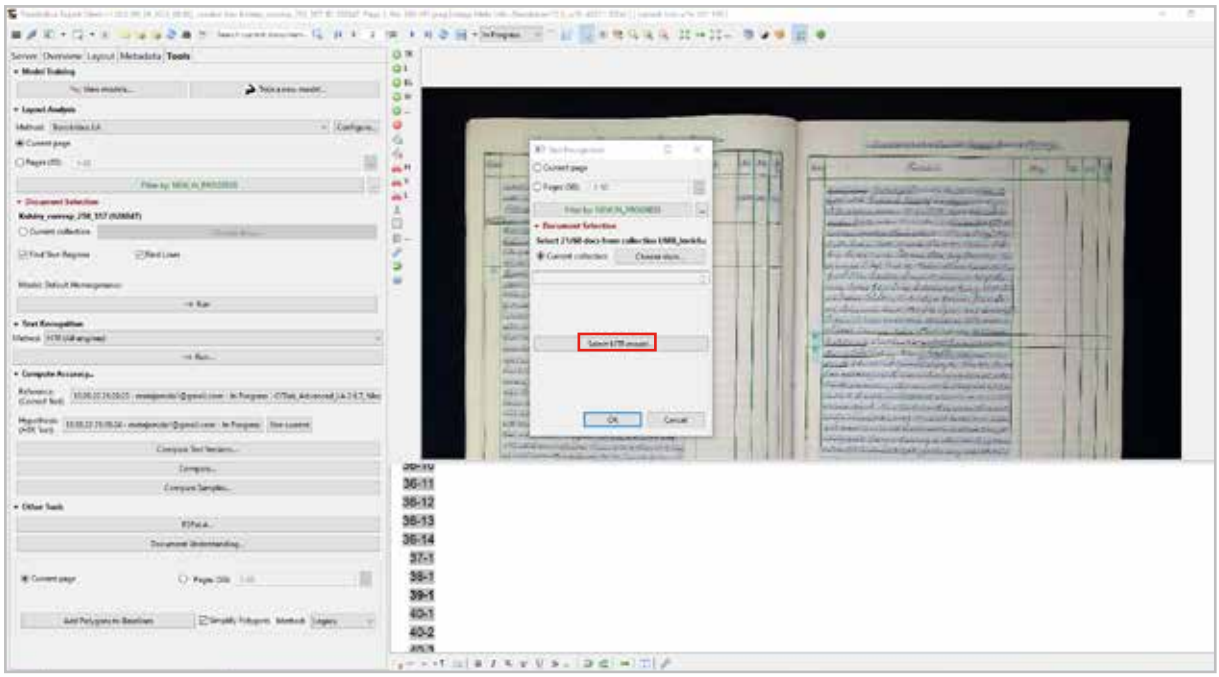


Obrázok 138 Výber snímok na automatickú transkripciu zo všetkých dokumentov v zbierke používateľa s aplikáciou filtra nastaveného na príznak snímky In progress

Výber modelu na automatickú transkripciu

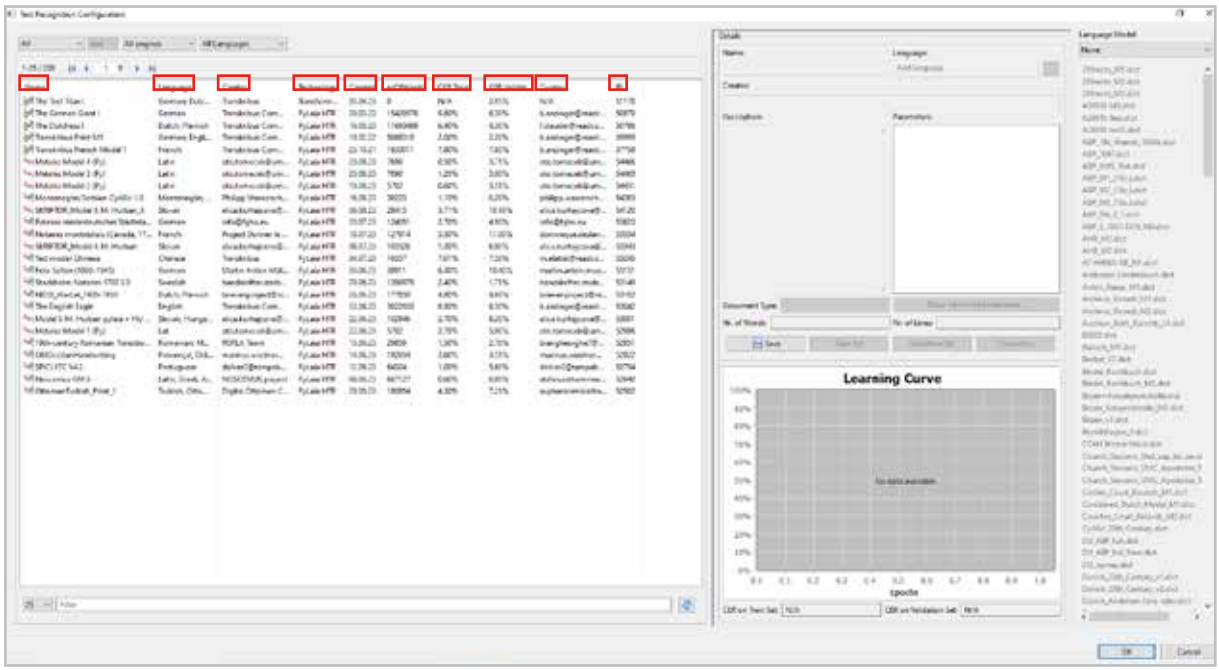
POZOR! V prípade, že ste si v Kroku 2 pri voľbe nástroja automatickej transkripcie zvolili možnosť Transkribus OCR (*Block-segmentation + Transkribus-Print-M1 model*) s predvolene nastaveným modelom na automatickú transkripciu tlačených dokumentov, nevykonávate ďalšie nastavenia špecifikované v tomto kroku a pokračujete Krokom 5.

Kliknutím na tlačidlo Vybrať HTR model (*Select HTR model…*) otvoríte samostatné okno, v ktorom vyberiete model, ktorý chcete aplikovať pri automatickej transkripcii vášho dokumentu.



Obrázok 139 Otvorenie okna pre výber modelu automatickej transkripcie

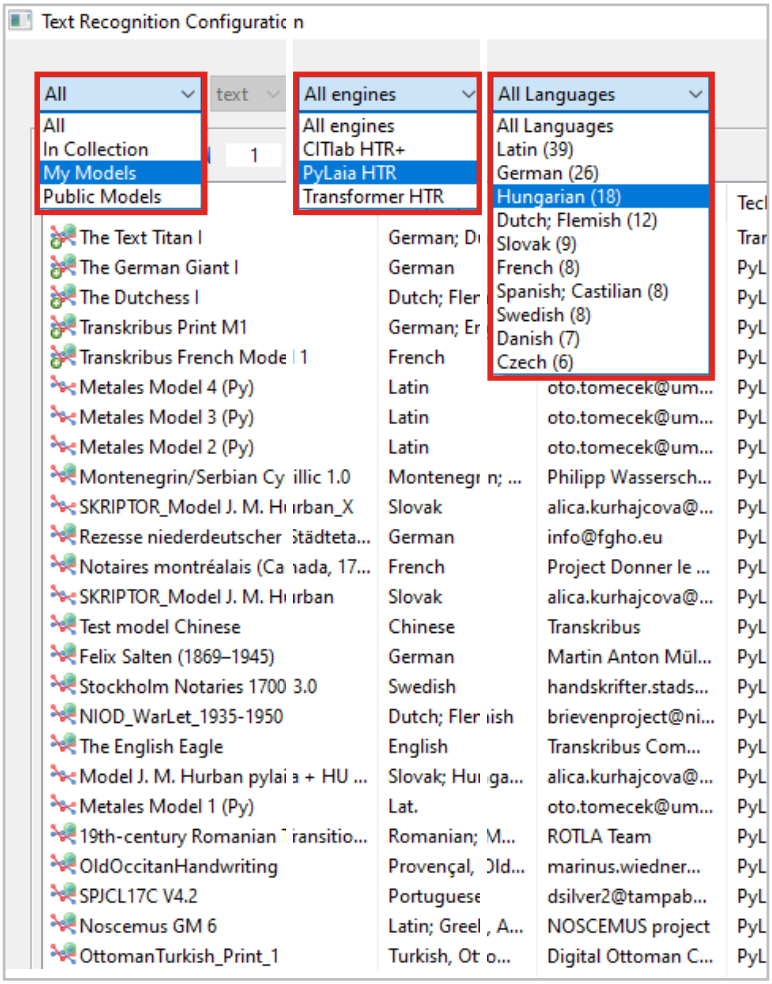
V ľavej časti okna na výber modelu vidíte zoznam všetkých dostupných modelov (verejných a vašich vlastných). Každý model má uvedené základné charakteristiky: jazyk; autora; použitú technológiu rozpoznania textu; dátum vytvorenia; počet slov v cvičnom súbore použitom k jeho vytrénovaniu; chybovosť transkripcie v cvičnom súbore; chybovosť transkripcie v overovacom súbore; osobu spravujúcu model; číselný identifikátor.



Obrázok 140 Zoznam dostupných modelov a ich charakteristík v okne pre voľbu modelu na automatickú transkripciu

Na lepšiu orientáciu môžete modely filtrovať cez ponuku v hornej lište.

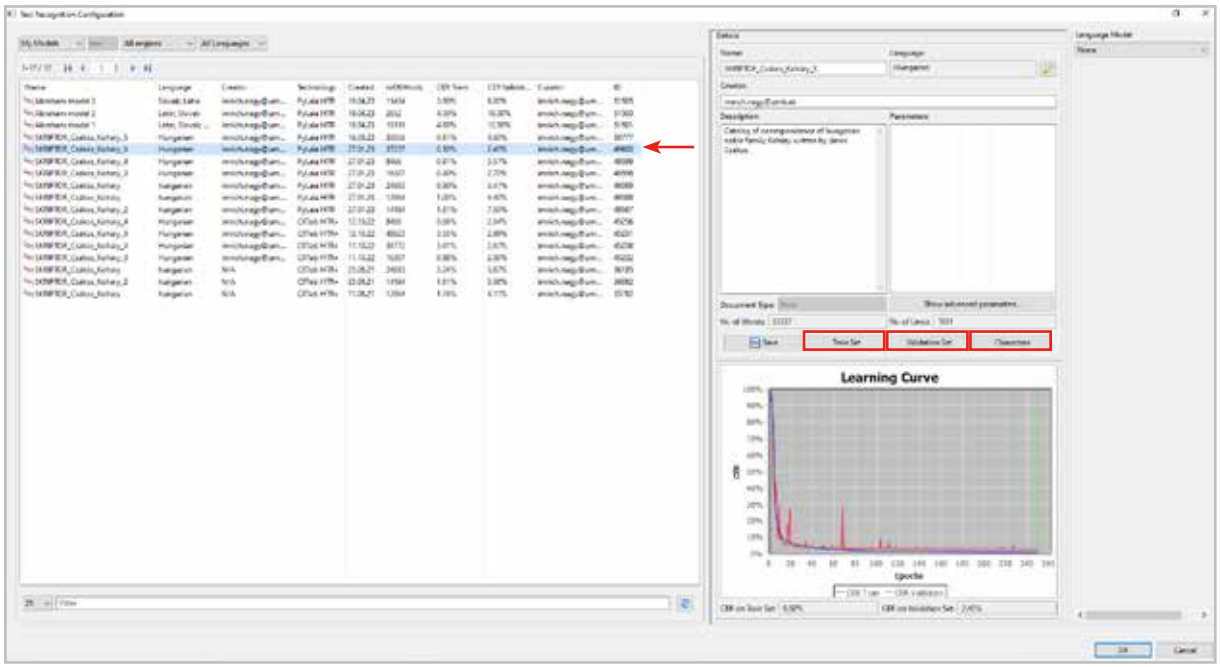
Poznámka: V súčasnosti platforma Transkribus podporuje iba technológiu PyLaia HTR.



Obrázok 141 Možnosti filtrovania modelov

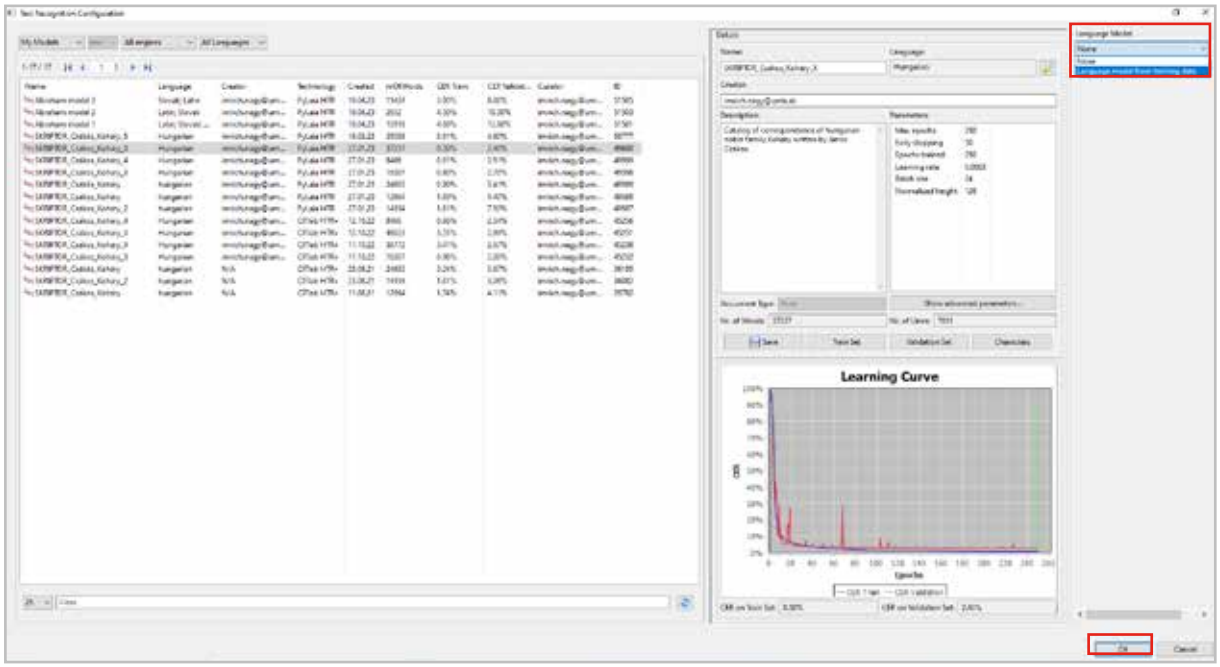
Keď si zo zoznamu prístupných modelov vyberiete vhodný model, v pravej časti obrazovky sa ukážu jeho detaily: názov modelu, jazyk modelu, autor modelu, stručný popis a dáta z jeho trénovania (počet slov, počet riadkov, grafický záznam priebehu chybovosti na cvičnom a overovacom súbore počas trénovania modelu a konečné údaje chybovosti CER na cvičnom a overovacom súbore). Ak kliknete na tlačidlá Cvičný súbor (*Training Set*) a Overovací súbor (*Validation Set*) máte možnosť overiť si aj vizuál jednotlivých strán (pôvodné digitalizáty) dokumentu použitého pri trénovaní modelu, resp. sadu znakov, ktoré boli použité pri ich prepise (tlačidlo Znaky (*Characters*).

Poznámka: Overiť si predlohu modelu má zmysel vtedy, ak chcete použiť model, ktorý ste nevytvárali, a teda vopred neviete, či bude zodpovedať (napr. typom písma) vášmu dokumentu. Pri verejných modeloch však náhľad pôvodných digitalizátov nemusí byť vždy dostupný.



Obrázok 142 Výber modelu a jeho charakteristiky v ľavej časti obrazovky s aktívnymi tlačidlami na zobrazenie pôvodných digitalizátov použitých do cvičného a overovacieho súboru modelu v pravej časti obrazovky

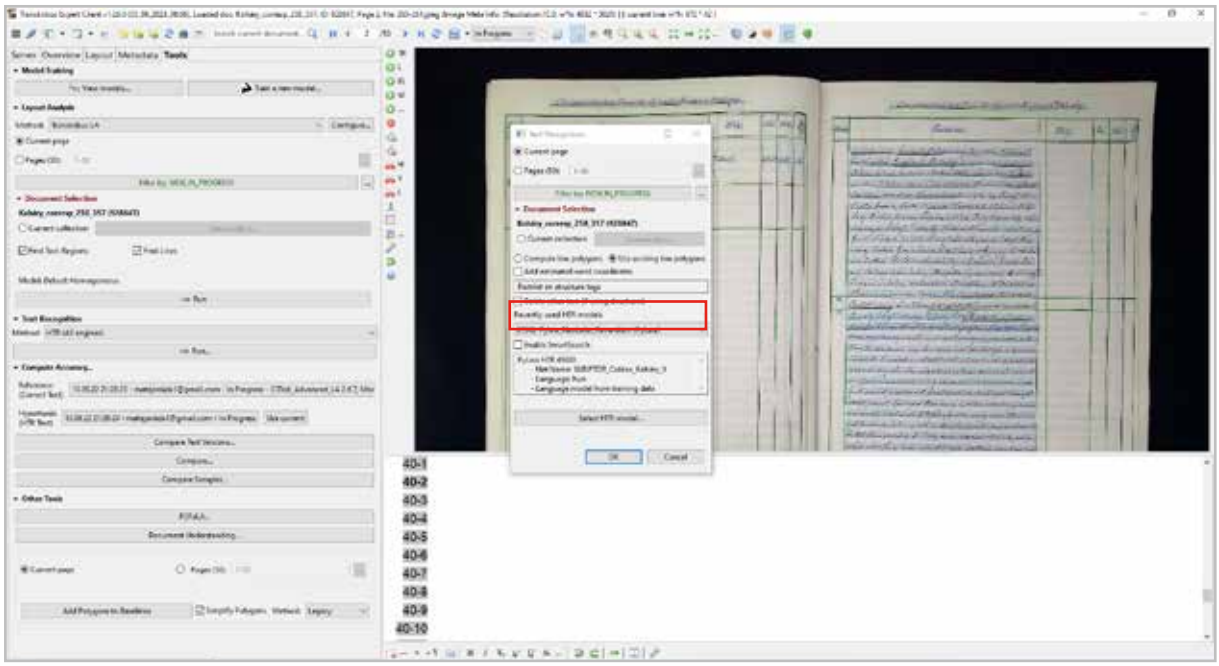
Vpravo hore sa nachádza možnosť pridať do predvolieb automatickej transkripcie aj **jazykový model**, ktorý sa automaticky vytvára pri trénovaní modelu. Pridanie jazykového modelu môže pomôcť najmä pri dokumentoch, kde sa niektoré výrazy často opakujú (napr. matriky, vizitačné protokoly, účtovné knihy a pod.). Po potvrdení výberu modelu kliknutím na tlačidlo OK vpravo dolu sa vrátite do dialógového okna Rozpoznávanie textu (*Text Recognition*).



Obrázok 143 Pridanie jazykového modelu do nastavení výberu modelu na automatickú transkripciu a potvrdenie výberu modelu

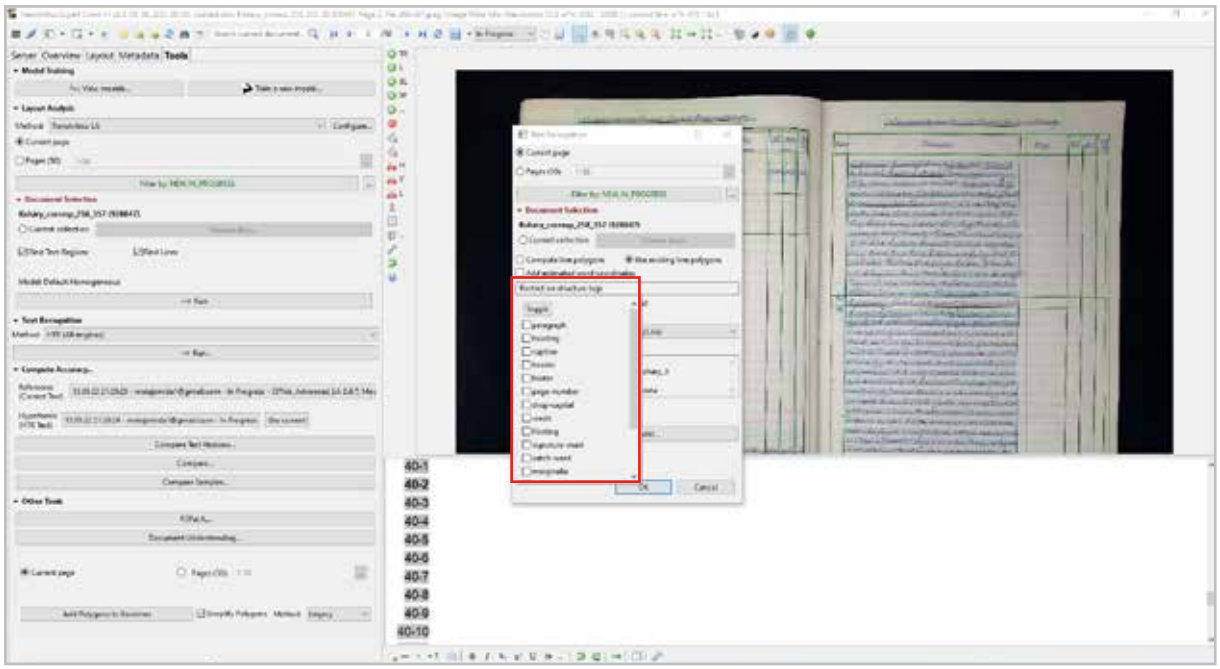
Záverečné nastavenie predvolieb automatickej transkripcie a jej spustenie
Po výbere modelu na automatickú transkripciu dokumentu sa po návrate do okna *Text Recognition* zobrazia možnosti pokročilých predvolieb:

- predvolená hodnota *Compute line polygons* – Transkribus automaticky nanovo určí hranice riadku,
- *Use existing line polygons* – **vyberte túto možnosť**, ak ste vo fáze segmentácie textu **manuálne upravovali/opravovali hranice riadku**,



Obrázok 144 Pokročilé nastavenia v dialógovom okne Text Recognition po potvrdení výberu modelu s označeným výberom možnosti Use existing line polygons

- *Add estimated words coordinates* – voľbou tejto možnosti sa v originálnom dokumente zvýraznia hranice slov určených a aplikovaných pri automatickej transkripci. Táto funkcionalita môže pomôcť pri dodatočnej kontrole a korekcii automatického prepisu.
- *Restrict on structure tags* – ak ste oblasti textu označovali tagmi (značkami), napr. marginália, hlavička, päta, číslo strany a podobne, môžete ich výberom jednotlivých tagov, ktoré sa rozbalia po kliknutí na toto tlačidlo, označiť, t. j. **obmedziť rozpoznávanie textu na označené tagy**,
- *Delete other text (if using structures)* – ak ste pri predchádzajúcej voľbe označili tagy, na ktoré sa má zamerať rozpoznávanie textu, voľbou tejto možnosti môžete text v ostatných oblastiach dokumentu odstrániť z automatického prepisu,



Obrázok 145 Ponuka tagov (značiek) na výber pri voľbe Restrict on structure tags

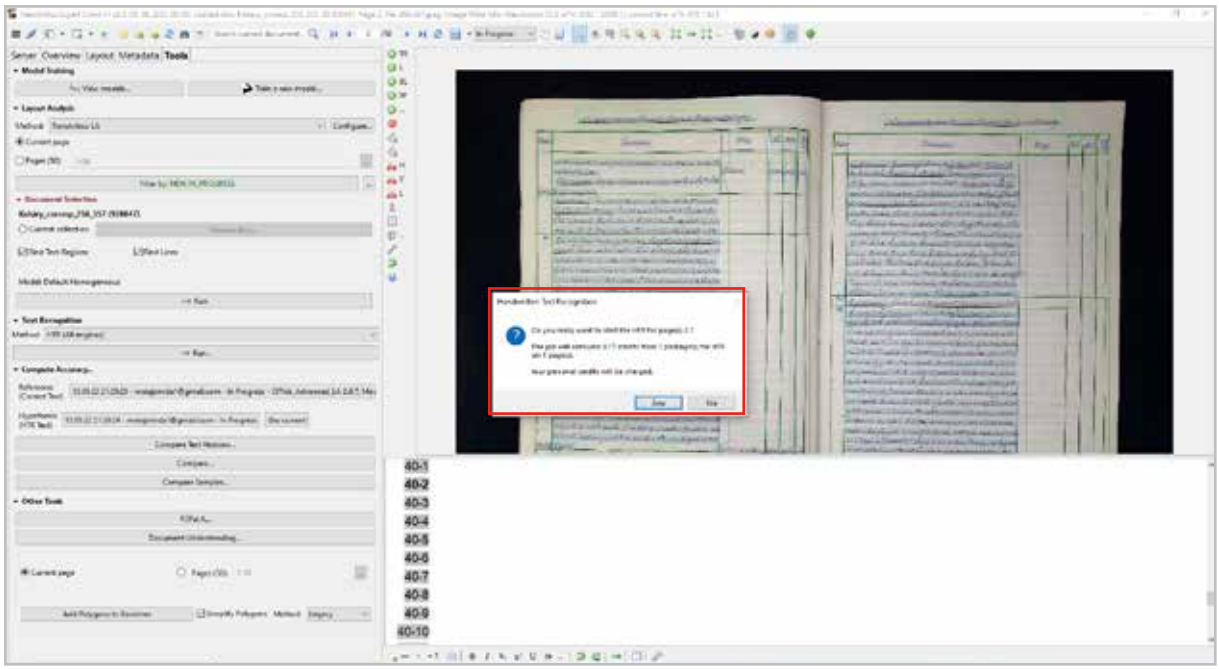
- *Recently used HTR models* – rozbaľovacie tlačidlo s ponukou nedávno použitých modelov ostatnými užívateľmi (odporúčame **nerozbaľovať**: náhodným kliknutím môžete vybrať model z tejto ponuky, čím sa prepíšu všetky predvoľby, ktoré ste predtým nastavili),
- *Enable SmartSearch* – príprava na využitie funkcie zdokonaleného vyhľadávania v automaticky transkribovanom dokumente (vo výsledkoch vyhľadávania sa objavia nielen výrazy identické s reťazcom zadaným pri vyhľadávaní, ale aj rôzne alternatívy príslušného slova. Predvolenie si Transkribus so zapnutou funkciou *SmartSearch* dokáže zapamätať, a potom aj vyhľadať až 100 variantov slova.

POZOR! Zapnutie tejto funkcie zvyšuje hardvérové nároky na výpočtový výkon aj úložný priestor voči serverom platformy Transkribus. Týmto odôvodňuje prevádzkovateľ platformy 50 % navýšenie poplatku za automatickú transkripciu každej jednej strany (snímky) rukopisného/tlačeného dokumentu so zapnutou funkciou *SmartSearch*.

Poznámka: Vzhľadom na túto cenovú politiku si zapnutie funkcie SmartSearch *treba vopred premyslieť. Na druhej strane, ak by ste chceli funkciu* SmartSearch *pridať dodtatočne, k už automaticky prepísanému dokumentu, musíte na celý dokument nanovo spustiť automatickú transkripciu so zapnutou funkciou* SmartSearch*, za čo vám bude účtovaný opätovne poplatok vo výške 1,5 násobku poplatku, ktorý ste už uhradili za automatickú transkripciu bez zapnutej funkcie* SmartSearch.

- v poslednom rámčeku sú pre kontrolu uvedené základné špecifikácie modelu (názov a jazyk modelu, použitie, resp. nepoužitie jazykového modelu), ktorý ste vybrali na automatickú transkripciu,

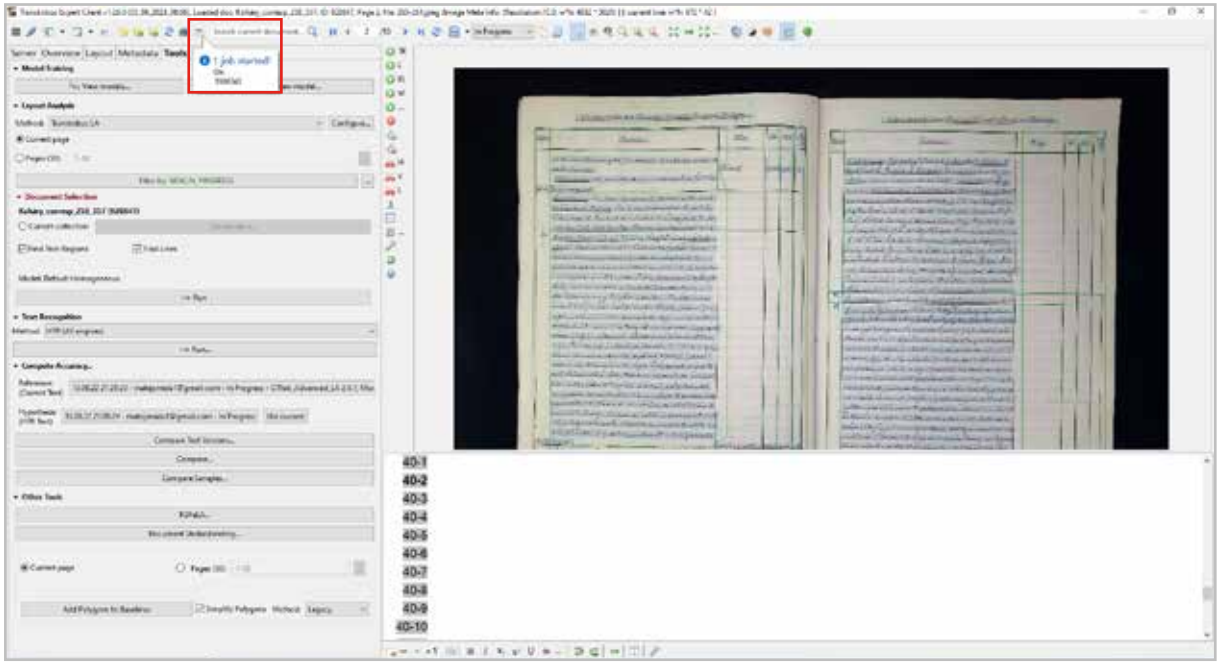
Kliknutím na potvrdzovacie tlačidlo OK spustíte automatickú transkripciu podľa zadefinovaných kritérií. Ešte pred spustením samotnej automatickej transkripcie sa otvorí okno s potvrdzujúcou otázkou a upozornením, že ide o spoplatnenú operáciu.



Obrázok 146 Okno s potvrdzujúcou otázkou na spustenie automatickej transkripcie s upozornením, že ide o spoplatnenú operáciu a s informáciou o výške poplatku za 1 stranu (snímku) digitalizátu

Výsledok automatickej transkripcie

O spustení automatickej transkripcie Vás informuje záložka pri ikone zoznamu úloh (*Jobs*).

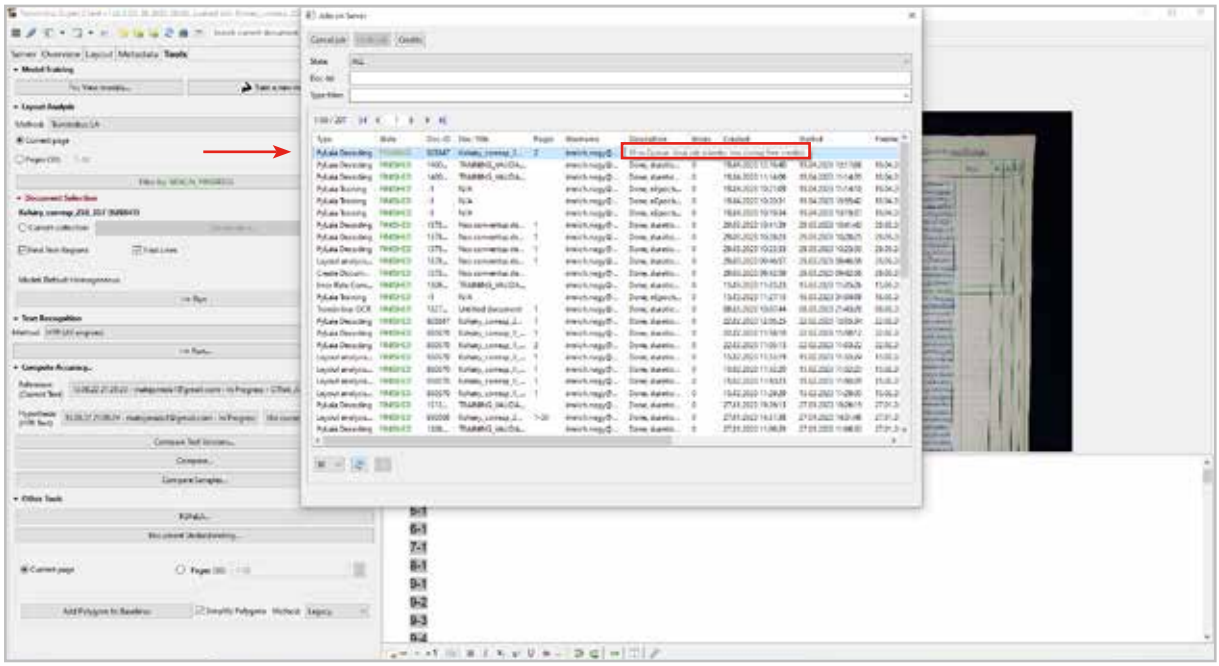


Obrázok 147 Záložka pri ikone zoznamu úloh informujúca o spustení operácie automatickej transkripcie

Po kliknutí na ikonu zoznamu úloh sa otvorí okno zoznamu spustených úloh na serveroch platformy Transkribus *Jobs on Server*. Operácia automatickej transkripcie, ktorú ste spustili, je uvedená v prvom riadku s príslušným statusom: *PENDING* / *RUNNING* / *FINISHED* a základným popisom.

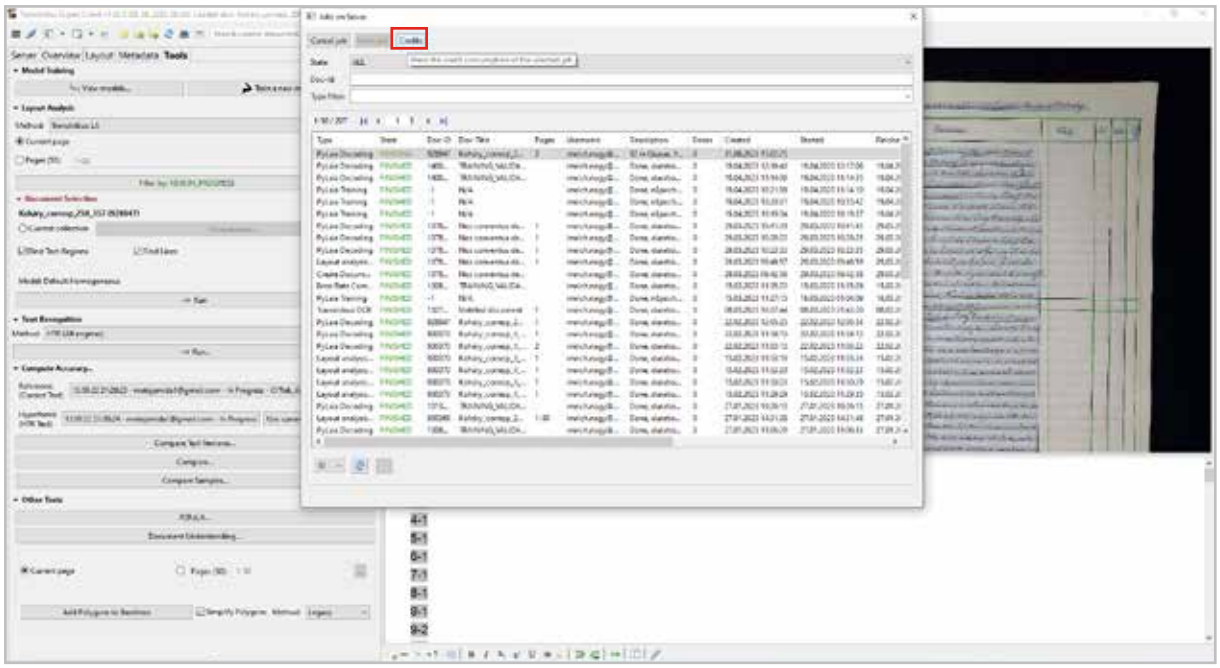
Vaša požiadavka na automatickú transkripciu sa zaradí do poradia podľa aktuálne spracovávaných požiadaviek na serveroch platformy Transkribus. Poradie vašej operácie zistíte, ak podržíte myšku nad riadkom spustenej operácie, resp. táto informácia sa nachádza aj v stĺpci *Description*.

Čakacia doba na výsledok závisí od poradia a náročnosti jednotlivých operácií a rádovo sa zvyčajne pohybuje v hodinách až dňoch, samotný proces automatickej transkripcie netrvá dlho – približne minútu na jednu snímku (v závislosti od dĺžky prepisovaného textu).

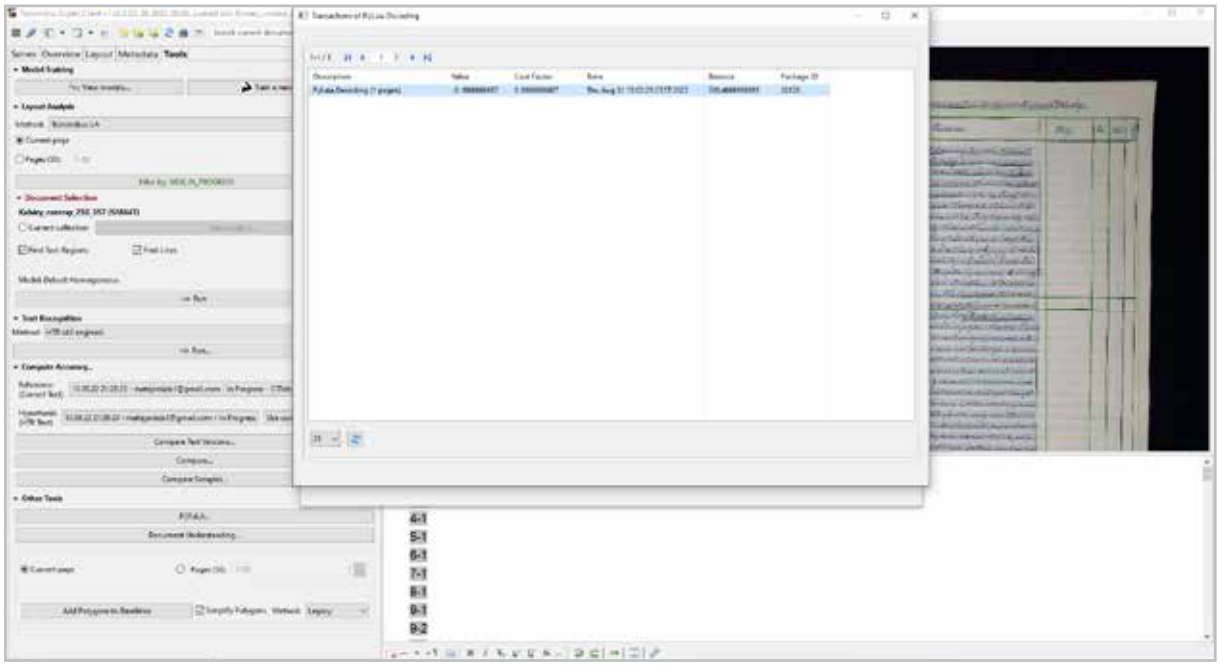


Obrázok 148 Okno so zoznamom bežiacich úloh na serveroch platformy Transkribus

Z okna zoznamu úloh si viete overiť aj cenu v kreditoch za zadanú požiadavku automatickej transkripcie po kliknutí na tlačidlo Kredity (*Credits*).

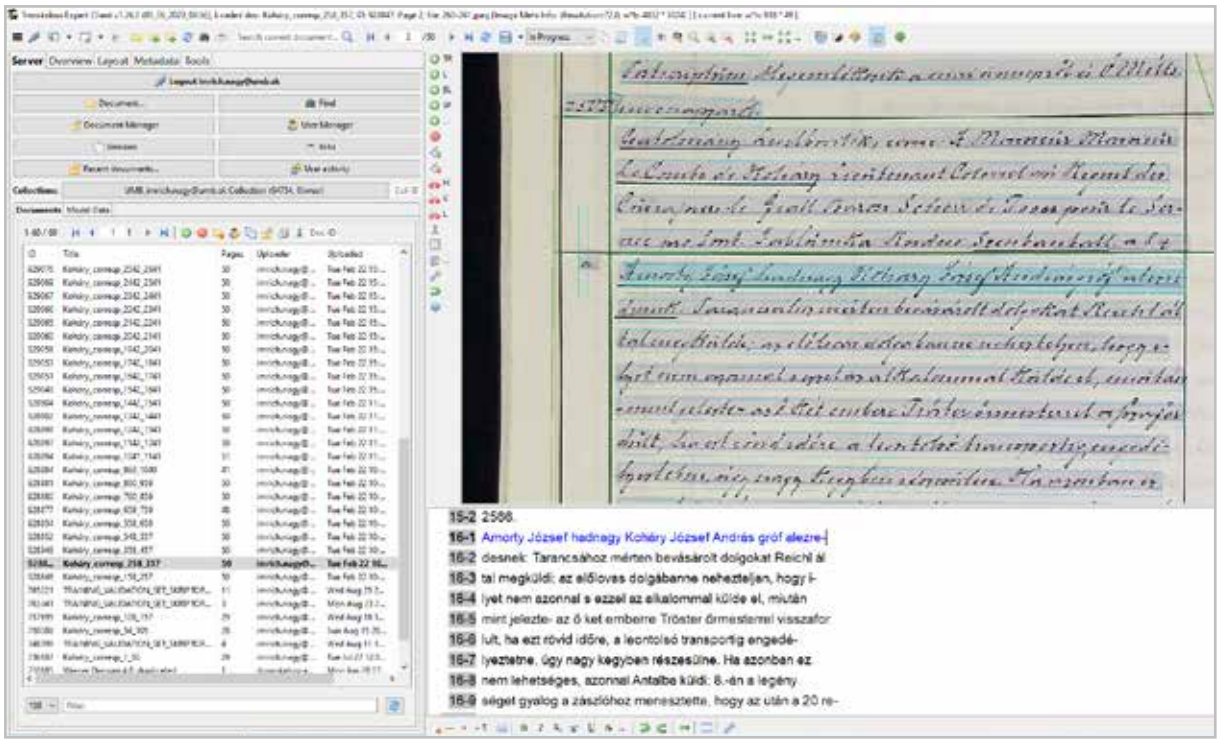


Obrázok 149 Tlačidlo Credits v okne zoznamu úloh na zobrazenie ceny (počtu kreditov) za vykonanie automatickej transkripcie



Obrázok 150 Stavový riadok v okne Transactions of PyLaia Decoding s informáciou o cene za automatickú transkripciu jednej snímky technológiou PyLaia HTR vo výške 0,17 kreditu

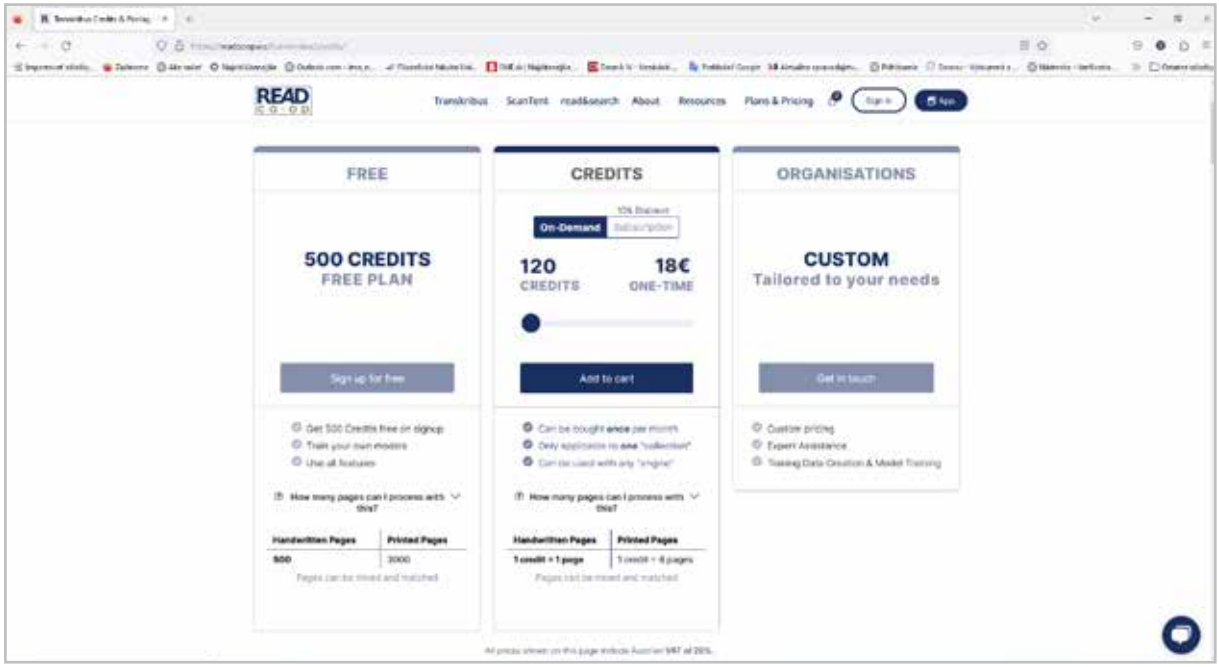
Po ukončení automatickej transkripcie sa pod snímkou príslušného digitalizátu zobrazí výsledok – prepis textu. Dvojitým kliknutím na riadok z prepisu sa farebne zvýrazní a priblíži príslušný riadok na digitalizáte, resp. vice versa, čo uľahčí kontrolu správnosti prepisu a prípadnú korekciu.



Obrázok 151 Výsledok automatickej transkripcie – prepis rukopisného textu so zvýrazneným riadkom na kontrolu správnosti

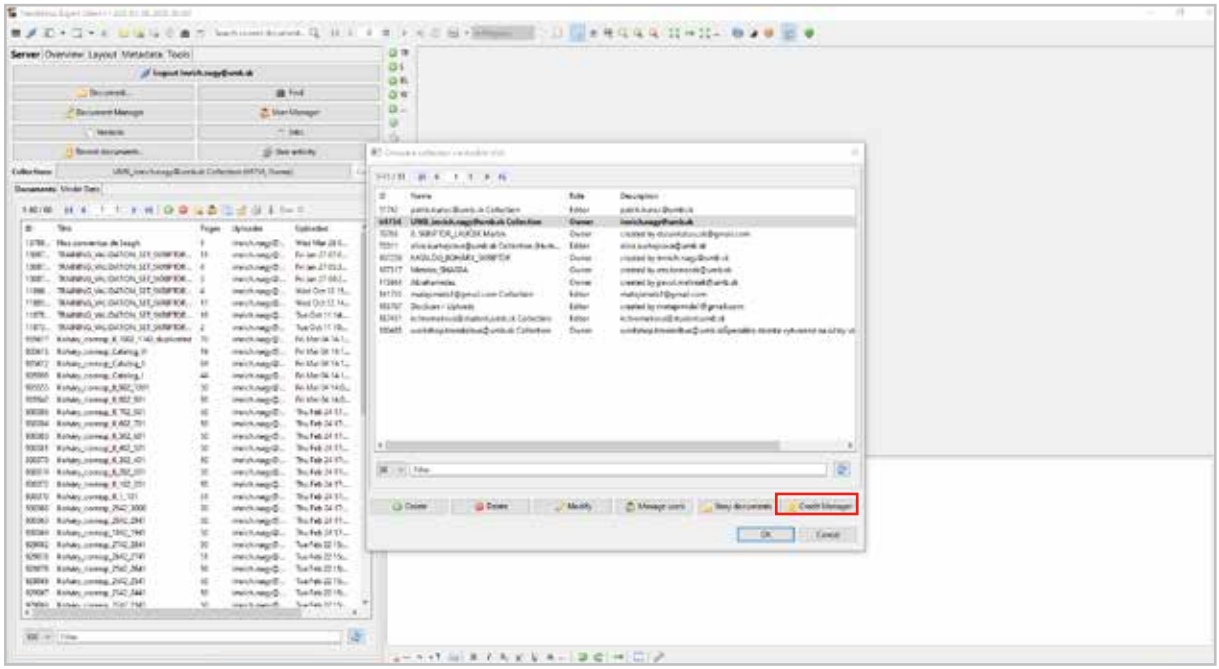
Kontrola kreditov a systém spoplatnenia automatickej transkripcie

Ako sa uvádzame vyššie, automatická transkripcia je spoplatnená formou kreditov. Okrem vstupného balíka 500 kreditov, ktoré získate zdarma pri registrácii na webovej stránke platformy Transkribus https://account.readcoop.eu, si môžete podľa potreby dokupovať kredity v e-shope platformy. Cenová politika sa priebežne môže meniť, momentálne sa základná cena odvíja od sumy 0,15 €/jeden kredit. Je to najvyššia suma bez množstevných alebo členských zliav, resp. zliav vyplývajúcich zo zakúpenia predplatného.



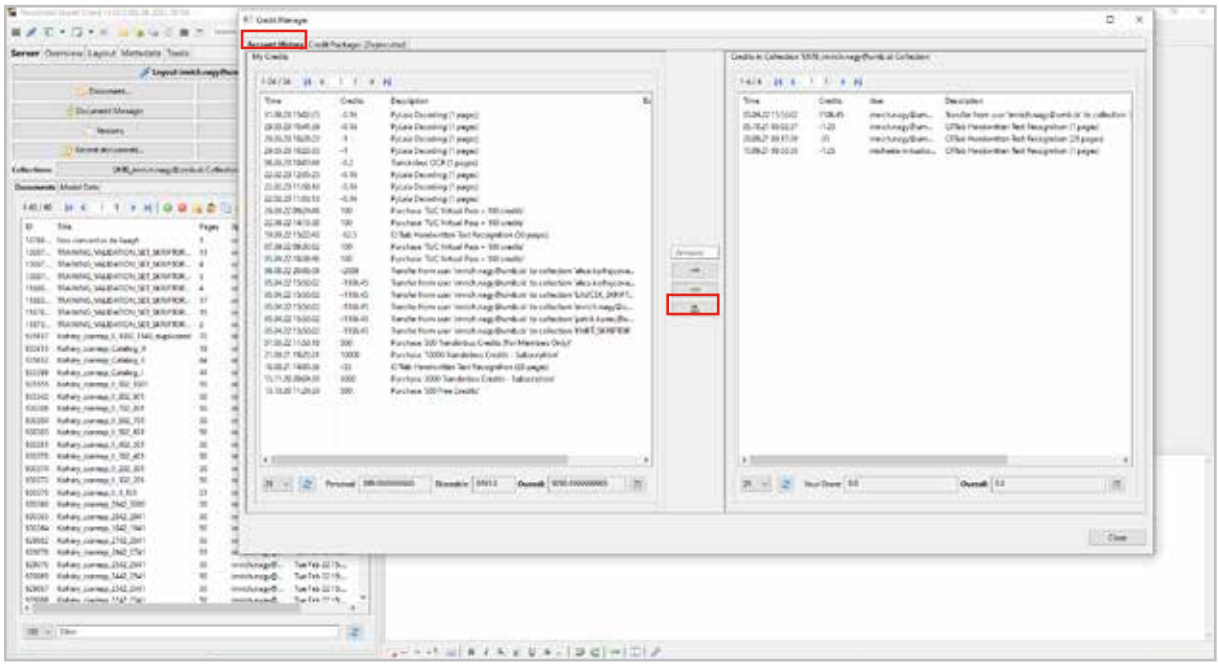
Obrázok 152 E-shop na nákup kreditov na webovej stránke https://readcoop.eu/transkribus/credits/

Počet disponibilných kreditov si môžete overiť priamo v aplikácii Transkribus expert klient, kde v hlavnom okne zvolíte svoju zbierku (stavový riadok *Collections*). Otvorí sa nové okno *Choose* a *collection via double click*, v ktorom treba kliknúť vpravo dolu na tlačidlo Správa kreditov (*Credit manager*).



Obrázok 153 Otvorenie správy kreditov pomocou tlačidla Credit Manager

V okne Správy kreditov (*Credit Manager*) na záložke História účtu (*Account History*) vidieť v ľavej časti históriu pohybov kreditov (nákup/pridelenie kreditov a ich spotreba na automatickú transkripciu). Po označení balíčka kreditov v ľavej časti, ktoré sú označené ako *Shareable* sa nachádza možnosť prerozdeliť ho medzi zbierky (kliknutím na tlačidlo uprostred), resp. účty iných osôb, ktoré sa zobrazia v pravej časti okna.

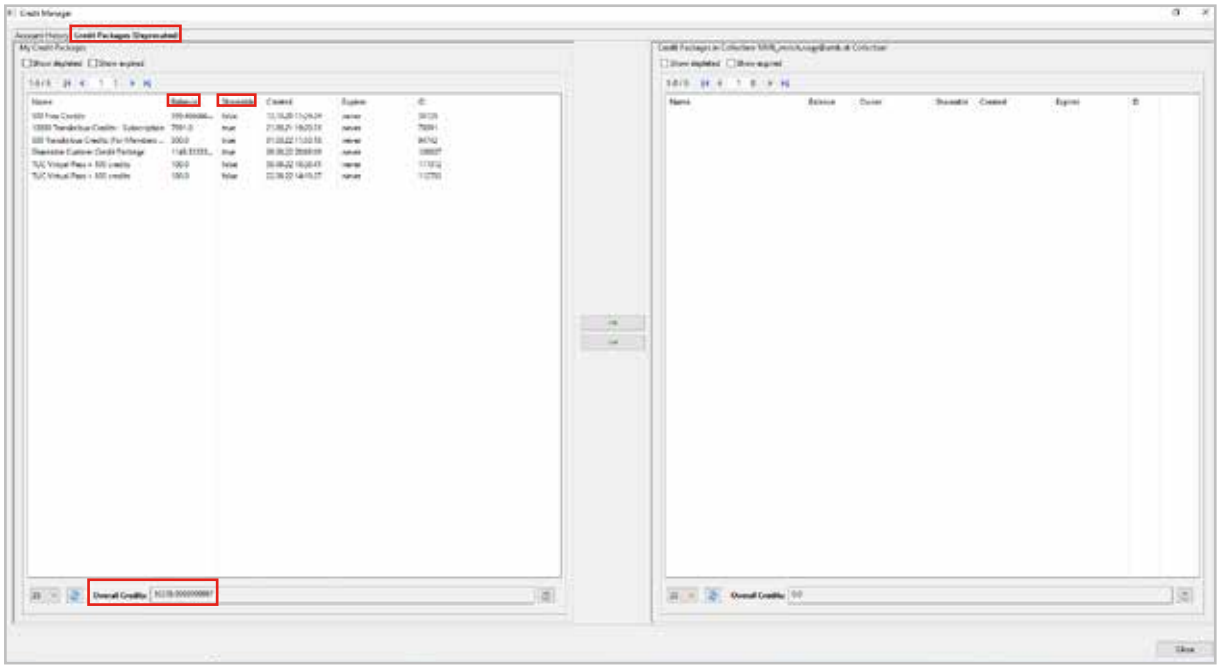


Obrázok 154 Okno správy kreditov Credit Manager – uprostred tlačidlo na presun kreditov na iný účet

Na záložke Balíky kreditov (*Credit Packages*) si môžete v ľavej časti overiť, o ktoré kredity sa môžete podeliť (v stĺpci *Shareable* je hodnota *True*) s iným kontom. V zásade ide o zakúpené kredity, teda kredity, ktoré vám boli pridelené pri registrácii konta, a ktoré nemôžete prideliť inému užívateľovi.

V stĺpci Zostatok (*Balance*) vidieť aktuálny stav jednotlivých balíčkov kreditov, ktoré máte na svojom konte. V spodnom riadku Celkové kredity (*Overall credits*) vidíte sumár všetkých dostupných kreditov, ktoré máte k dispozícii.

Vyčerpané balíčky kreditov (*Depleted*) sa zobrazujú v pravej časti okna.



Obrázok 155 Záložka Credit Packages, na ktorej si môžete overiť stav, platnosť a možnosť rozdelenia kreditov

7 Možnosti práce s textom po automatickej transkripcii

Kapitola uvádza možnosti práce s prepísaným textom, ktorá ho zmení na dátovú základňu a export požadovaných obrazových či textových informácií, s ktorým chcete pracovať už v inom formáte alebo inom programe.

Text získaný po automatickej transkripcii a jej kontrole môžete obohatiť o dodatočné informácie. Spočíva vo vyčlenení významných údajov v rámci textu podľa nastavených kritérií. Uskutočňuje sa jeho označením zodpovedajúcimi tagmi (značkami).

Rozlišujeme dva základné typy tagov:

1. **Textové tagy**, ktoré definujú pojmy a frázy v texte a slúžia na označenie na úrovni oblasti, riadku, slova alebo aj jednotlivých znakov. Úpravy urobíte pravým kliknutím na prepísaný text v textovom editore.
2. **Štrukturálne tagy**, ktoré definujú štruktúru dokumentu a sú založené na oblastiach textu a riadkov. Úpravy urobíte pravým kliknutím na digitalizovanú snímku v obrazovej časti pracovnej plochy.

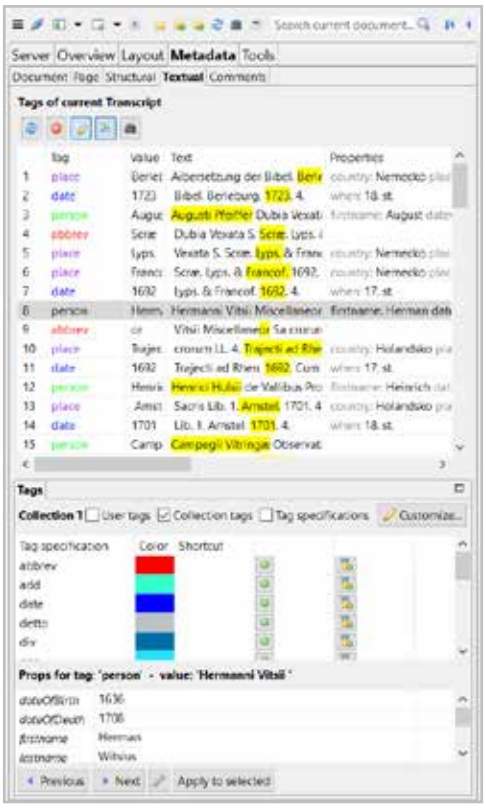
7.1 Textové tagy

Prepísané texty môžete obohatiť o textové tagy, ktoré bližšie charakterizujú zvolený výraz. Platforma ponúka preddefinované tagy označené kurzívou, s ktorými môžete pracovať ihneď alebo si môžete vytvoriť vlastné tagy. Práca s tagmi je možná až po priradení konkrétneho tagu k požadovanému výrazu.

Rozlišujeme:

1. autoritatívne tagy (napr. osobné meno, geografické miesto, dátum, inštitúcia, abstraktná identita),
2. ostatné textové tagy (napr. skratky, nečitateľné výrazy, vymazaný text, začiernený text),
3. vlastné tagy.

Funkcie na značenie textových tagov nájdete kliknutím na záložku Metadáta (*Metadata*) a následne na záložku Textové tagy (*Textual*) v ľavej hornej časti hlavnej pracovnej plochy. Záložka je prepojená s pracovnou plochou a textovým editorom.



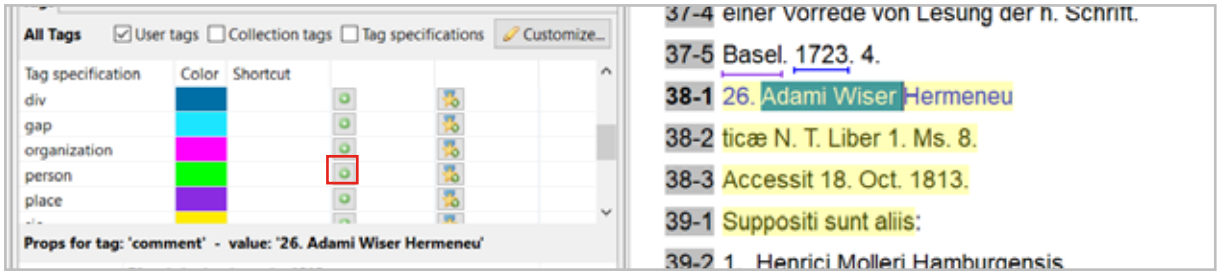
Obrázok 156 Štruktúra záložky Textové tagy (Textual)

7.1.1 Priradenie textového tagu

Textové tagy môžete použiť na výrazy na úrovni oblasti, riadku, slova alebo aj jednotlivých znakov. Označujte však len nevyhnutné časti textu, ktoré majú byť vyhľadateľné. Každý tag sa používa samostatne na zvolený výraz, ale v prípade potreby je možné k rovnakému výrazu priradiť aj viacero tagov.

Možnosti priradenia textového tagu k požadovanému výrazu v textovom editore:

1. Zvýraznite text v textovom editore a kliknite na zelené tlačidlo pri tagu, ktorý chcete použiť.

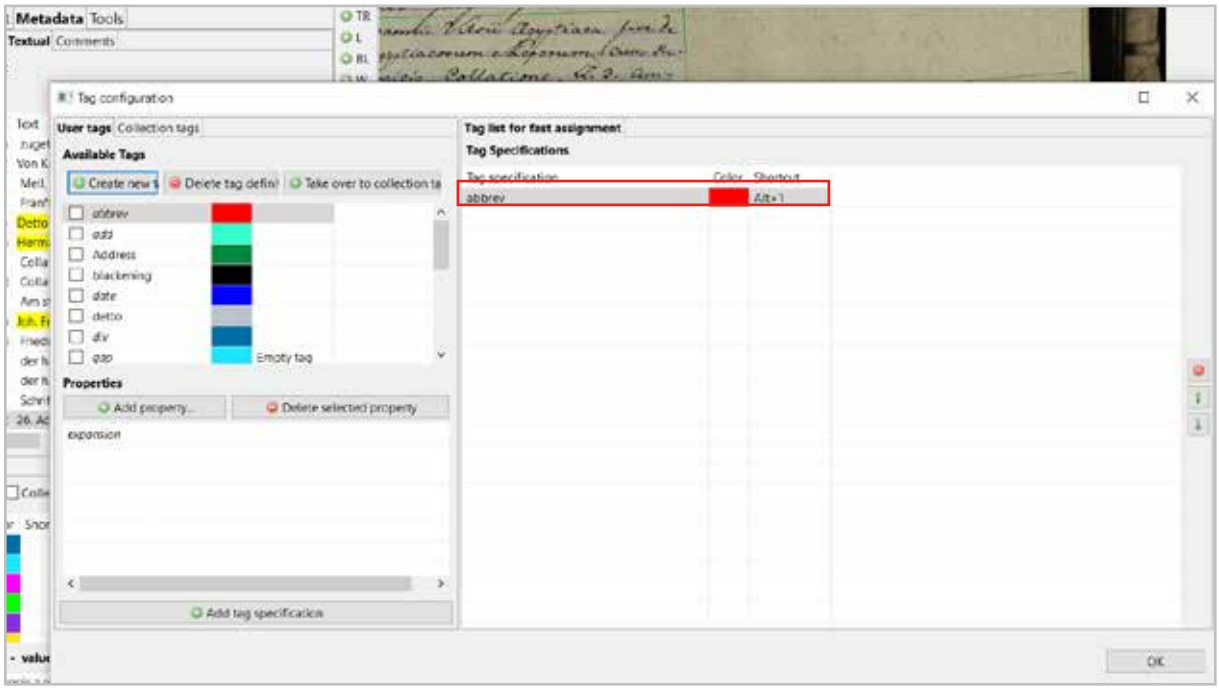


Obrázok 157 Označenie tagu pomocou zeleného tlačidla

2. Zvýraznite text a po kliknutí pravým tlačidlom myši vyberte možnosť Všetky tagy (*All tags*). Zobrazia sa definované tagy.

Obrázok 158 Označenie tagu pomocou myši

3. Častejšie používané tagy môžete priradiť aj prostredníctvom pridelenej skratky. Na zjednodušenie postupu kliknite na záložke Textové tagy (*Textual*) na tlačidlo Prispôsobiť (*Customize*). Po zobrazení okna Konfigurácia tagu (*Tag configuration*) označte požadovaný tag a kliknite na tlačidlo Pridať špecifikáciu tagu (*Add tag specification*). Do stĺpca Skratka (*Shortcut*) pridajte vhodnú skratku. Následne označte požadovaný text v textovom editore a zvoľte pridelenú skratku.

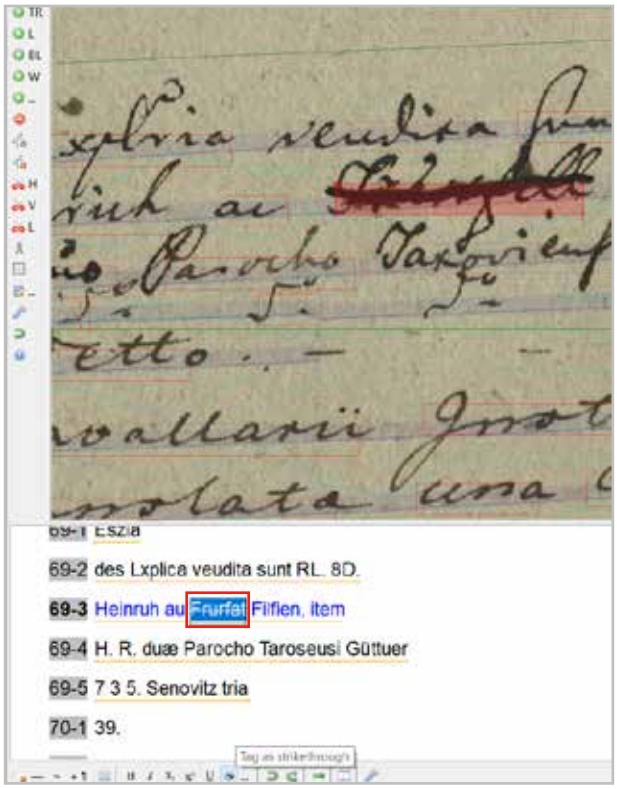


Obrázok 159 Označenie tagu pomocou skratky

7.1.2 Ostatné textové tagy

Patria do skupiny preddefinovaných tagov, ktoré majú stanovené pravidlá na použitie. Využívajú sa pri označovaní úprav textu daných dokumentov alebo dodatočných. V prameňoch sa z rôznych dôvodov vyskytuje aj nečitateľný text, ktorý sa nedá presne a dôveryhodne prepísať alebo nie je čitateľný.

Príklad 1 Ak je pôvodný text prešKrtnutý, ale stále čitateľný, prepíšte ho čo najvernejšie a dodatočne ho označte ako prečiarknutý prostredníctvom tlačidla Prečiarknutie (S*trikethrough*) na spodnej lište ponuky textového editora.

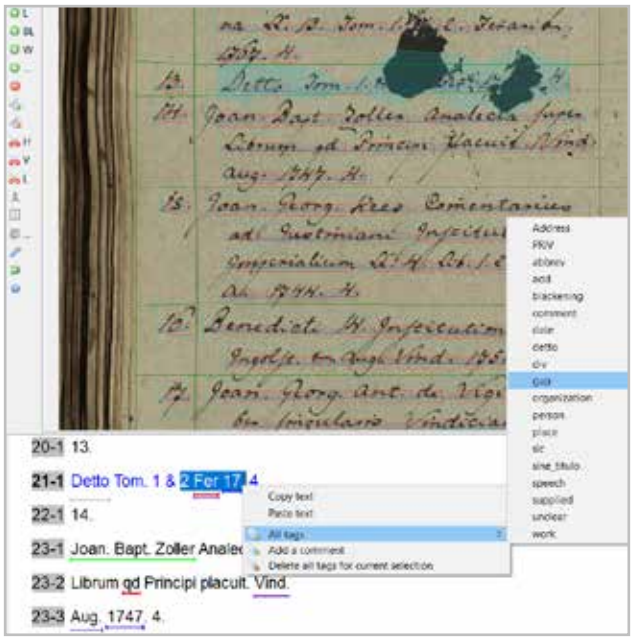


Obrázok 160 Použitie tlačidla Prečiarknutie (Strikethrough)

Príklad 2 Ak si nie ste istí správnosťou prepisu, prepísaný text označte tagom Nejasný (*Unclear*), aby ste sa ním mohli zaoberať neskôr. Riadky s takto označeným výrazom nie sú zahrnuté do trénovania modelu.

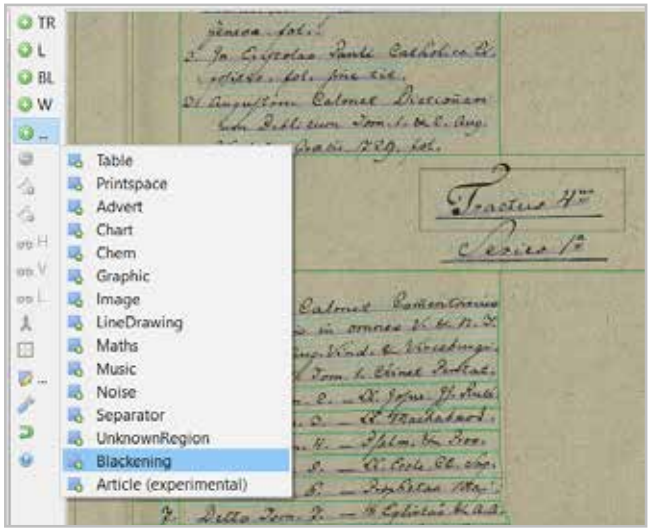
Obrázok 161 Použitie tagu Nejasný (Unclear)

Príklad 3 Ak je text úplne nečitateľný, také miesto označte tagom Medzera (*Gap*).

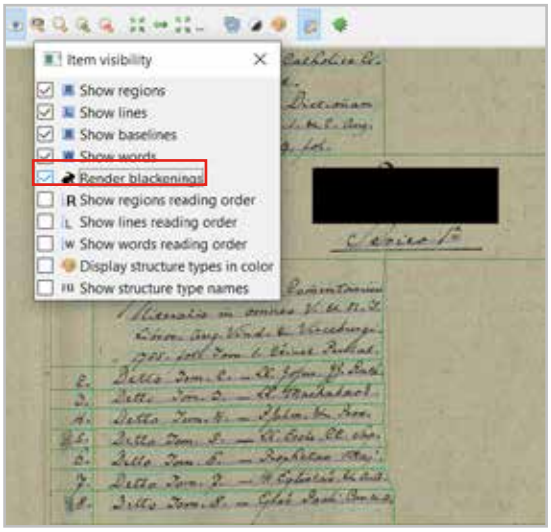


Obrázok 162 Použitie tagu Medzera (Gap)

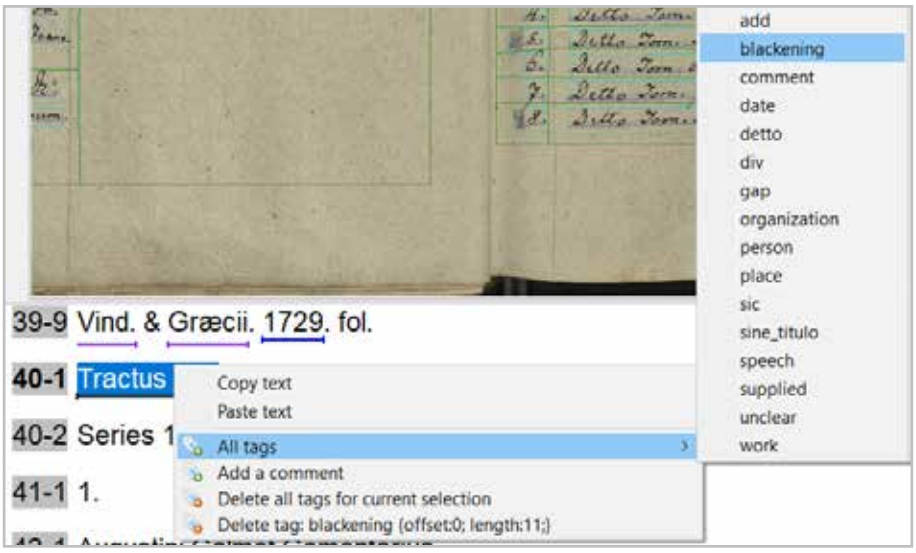
Príklad 4 Ak potrebujete prekryť čitateľné citlivé informácie, použite tag Začiernenie (*Blackening*). Používa sa v spojení s oblasťou „sčernenie“, ktorá sa pridáva pomocou segmentačných nástrojov. Použite rozbaľovaciu ponuku na tlačidle segmentačného prvku „+…“ na bočnej lište ponuky Plátno (*Canvas*) a vyberte možnosť Začiernenie (*Blackening*). Rámčekom označte slovo alebo text, ktorý chcete skryť. Potom kliknite na tlačidlo Viditeľnosť položky (*Item visibility*) na hornej lište ponuky Plátno (*Canvas*) a označte možnosť Vykresliť sčernenie (*Render blackenings*) na zobrazenie začiernených častí na snímke. Nakoniec kliknite pravým tlačidlom myši na zodpovedajúci výraz v poli textového editora a z voľby Všetky tagy (*All tags*) vyberte tag Začiernenie (*Blackening*). Pri exporte dokumentu vyberte možnosť Vykonať začiernenie (*Do blackening*) a text sa nahradí takto: […]. Informácie za začiernenou časťou sú zachované len v súboroch METS a TEI, v iných formátoch súborov je text úplne prekrytý.



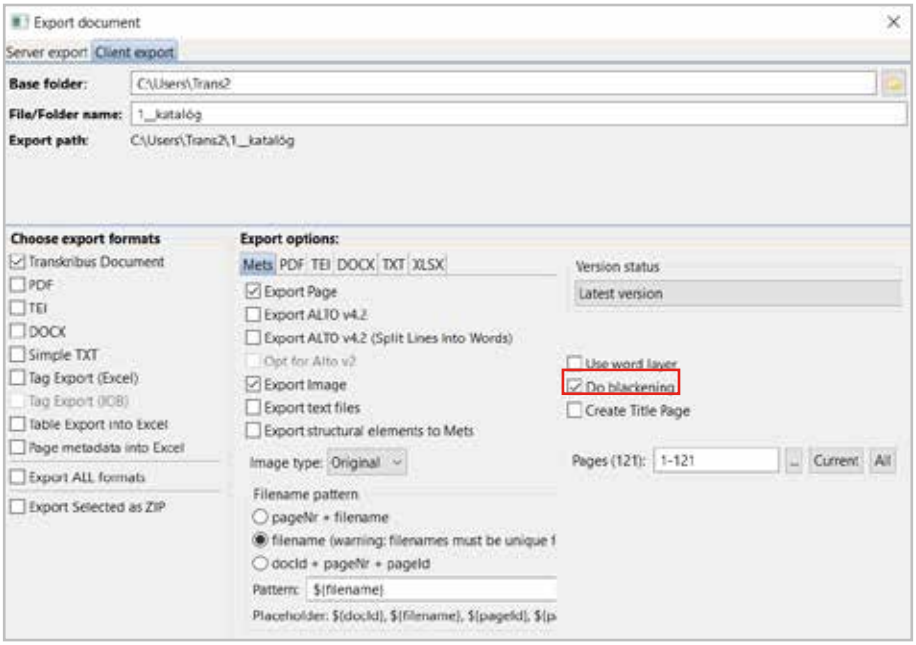
Obrázok 163 Označenie prekrytia



Obrázok 164 Zviditeľnenie prekrytia



Obrázok 165 Označenie tagom Začiernenie (Blackening)



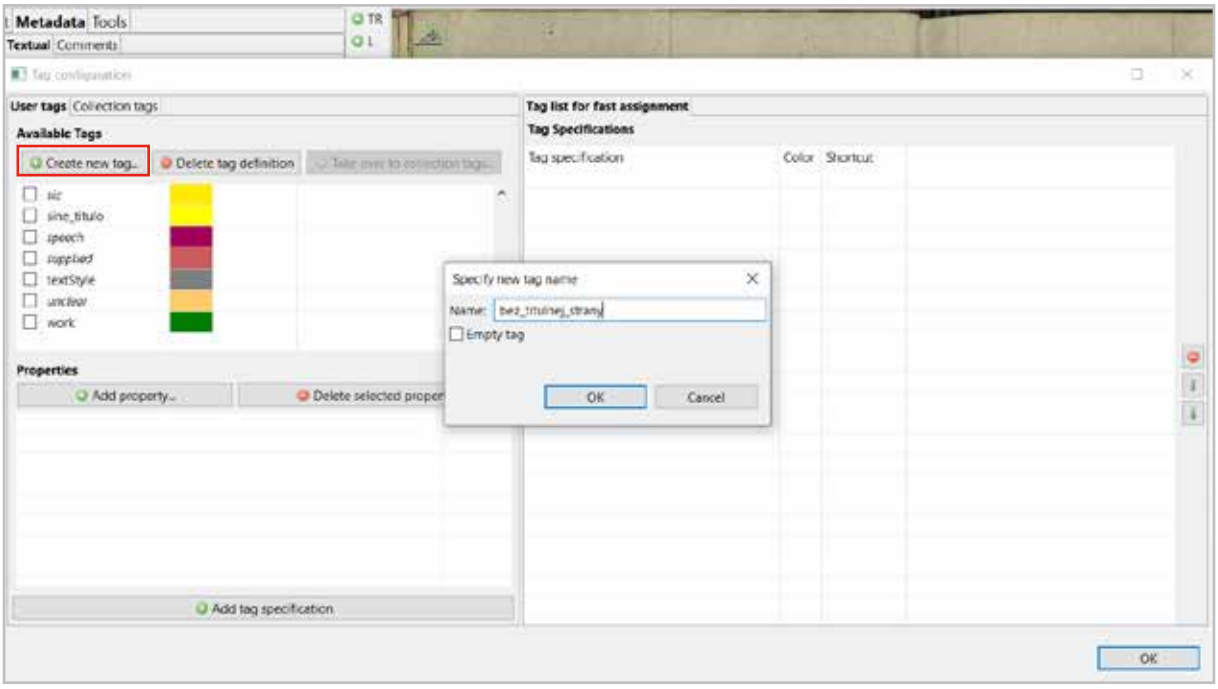
Obrázok 166 Označenie prekrytia v okne exportu

Môžete tiež pridať alternatívy a návrhy k textu alebo dôvody pre nečitateľný text voľbou Vlastnosti (*Properties*) cez okno Konfigurácia tagu (*Tag configuration*).

V niektorých prípadoch sa dá nečitateľný znak alebo znaky uhádnuť a tak sa dajú jednoducho prepísať. Namiesto pridania obvykle používaných hranatých zátvoriek doplnený text označte tagom Nahradené (*Supplied*).

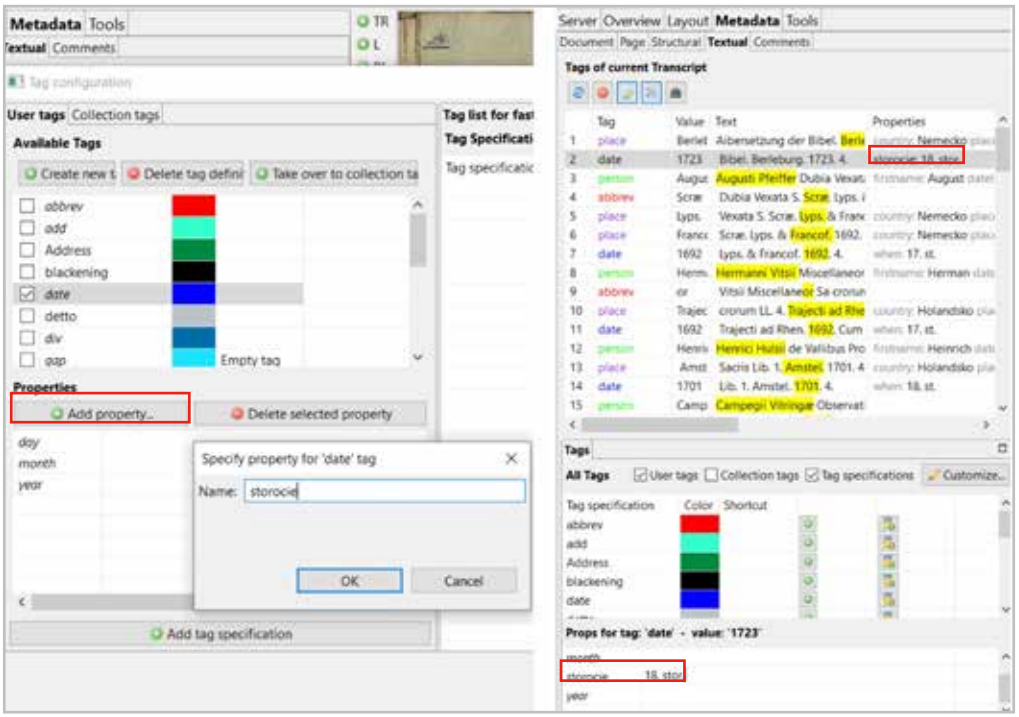
7.1.3 Vytvorenie nového textového tagu

Okrem preddefinovaných, ihneď dostupných tagov môžete používať aj ľubovoľné vlastné tagy. Vlastný tag vytvoríte kliknutím na tlačidlo Prispôsobiť (*Customize*) v strednej časti záložky Textové tagy (*Textual*). Po otvorení okna Konfigurácia tagu (*Tag configuration*) kliknite na tlačidlo Vytvoriť nový tag (*Create new tag*) a pomenujte ho.



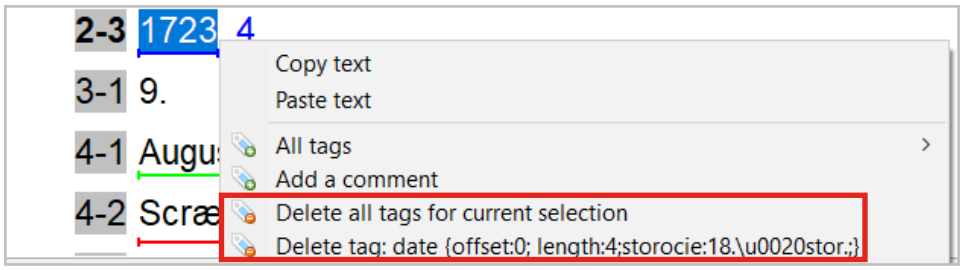
Obrázok 167 Vytvorenie nového tagu

V okne Konfigurácia tagu (*Tag configuration*) môžete nastaviť aj vlastnosti tagu, ktoré popisujú podrobnosti o konkrétnom výraze. Označte požadovaný tag a kliknutím na tlačidlo Pridať vlastnosť (*Add property*) sa otvorí okno, do ktorého zapíšete špecifickú vlastnosť tagu. Podrobnosti o výraze následne vyplňte pri jeho označení tagom.



Obrázok 168 Vytvorenie vlastnosti storočie pre tag Dátum (Date) a jeho použitie v texte

Ak ste výraz/text označili omylom, môžete tento krok opraviť. Opätovne zvýraznite text a kliknutím pravého tlačidla myši stlačte tlačidlo Odstrániť (*Delete*). Na výber máte dve možnosti, odstrániť iba jeden tag alebo všetky tagy z aktuálneho výberu.



Obrázok 169 Zmazanie označenia

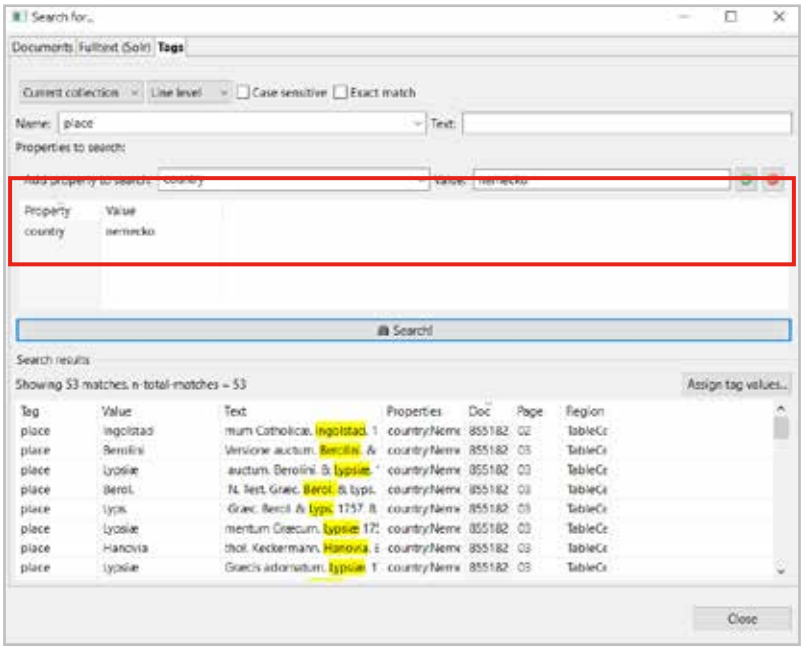
7.1.4 Vyhľadávanie textových tagov

Použité textové tagy a k nim priradené výrazy sa dajú vyhľadávať. Môžete s nimi pracovať už aj v expert klientovi, čo je užitočné napríklad ako pomôcka pri ďalšej práci s prameňom. Jednotlivé kroky sa uskutočňujú po zakliknutí tlačidla ďalekohľadu. Táto ikona je dostupná na viacerých miestach pracovnej plochy, napr. na hornej lište a na záložke *Server*, ale jej prednastavenie na vyhľadávanie tagov Tagy (*Tags*) je dostupné na záložkach Metadáta (*Metadata*) a Textové tagy (*Textual*).

Postup krokov na zadefinovanie požiadavky na vyhľadávanie:

1. vyberte miesto prehľadávania (napr. aktuálna zbierka, dokument, stránka),
2. zvoľte typ vyhľadávania, jednoduché (názov tagu, označené výrazy) alebo rozšírené (vlastnosti tagov a ich hodnoty),
3. hľadanie môžete obmedziť aj voľbou Rozlišovať malé a veľké písmená (*Case sensitive*).

Výsledky vyhľadávania sa zobrazia v spodnej časti okna po kliknutí na tlačidlo Hľadať (*Search*). Zobrazí sa prehľad informácií o tagu a jeho vlastnostiach (označený výraz, časť textu, číslo dokumentu a strany). Po dvojitom kliknutí na konkrétny riadok výsledku vyhľadávania sa na pozadí okna textového editora prekliknete na požadovanú stranu a výraz.

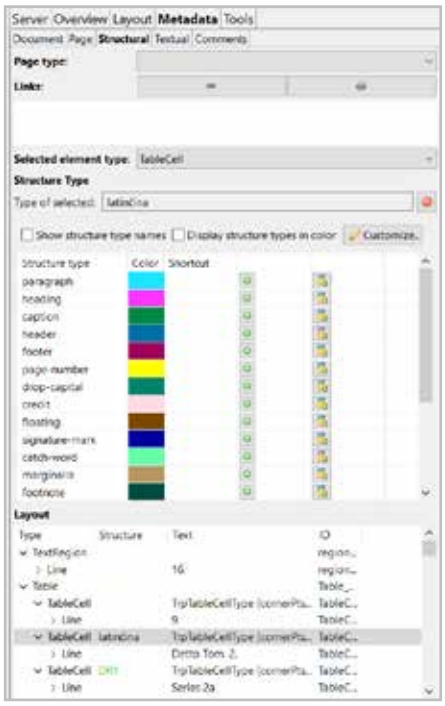


Obrázok 170 Výstup z rozšíreného vyhľadávania pre tag Miesto (Place)

7.2 Štrukturálne tagy

Prepísané texty môžete obohatiť aj o štrukturálne tagy (napr. odsek, nadpis, čísla strán, marginálie), ktoré umožňujú definovať štruktúru dokumentov. Je to doplnková funkcia, ktorú môžete využiť na označenie sekcií, ktoré vás zaujímajú (napr. vyčlenenie rôznych typov rukopisu v dokumente). Nie je potrebné označovať každý prvok dokumentu.

Nástroje na značenie štrukturálnych tagov nájdete na záložke Metadáta (*Metadata*) a následne na záložke Štrukturálne tagy (*Structural*).



← zoznam všetkých tagov

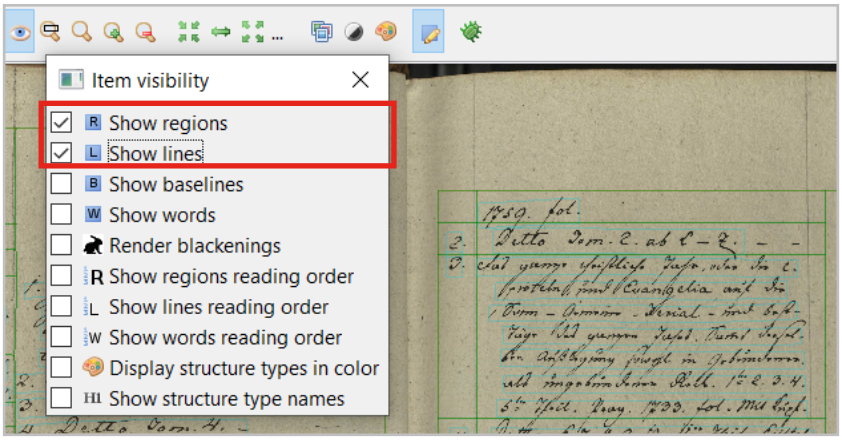
← prehľad použitých tagov

Obrázok 171 Štruktúra záložky Štrukturálne tagy (Structural)

7.2.1 Priradenie štrukturálneho tagu

Štrukturálne tagy sa priraďujú k oblastiam textu a oblastiam riadkov. Môžete označiť niekoľko oblastí naraz podržaním tlačidla CTRL na klávesnici a následne kliknutím na dokument.

Najprv kliknite na tlačidlo Viditeľnosť položky (*Item visibility*) v pravej časti hornej lišty hlavného menu, aby ste označili za viditeľné oblasti textu a oblasti riadkov.



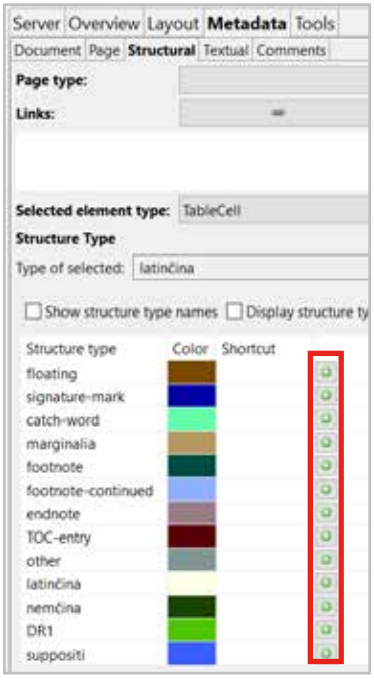
Obrázok 172 Potvrdenie viditeľnosti oblastí

Možnosti priradenia štrukturálneho tagu v okne snímky:

1. kliknite pravým tlačidlom myši na vybraný tvar a zvoľte požadovaný štrukturálny tag v časti Priradiť typ štruktúry (*Assign structure type*) z vyrolovaných možností,

Obrázok 173 Priradenie štrukturálneho tagu pomocou myši

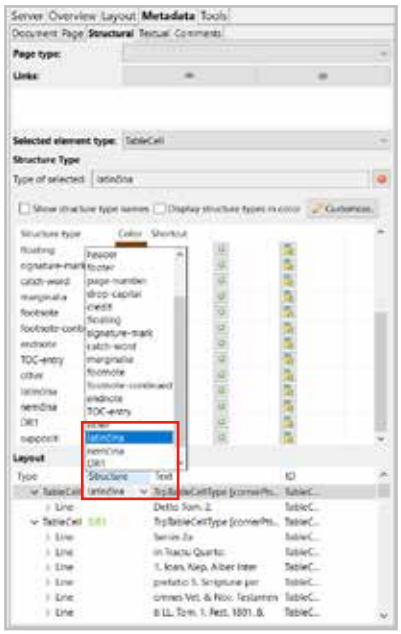
2. kliknite na zelené tlačidlo + pri požadovanom tagu na záložke Štrukturálne tagy (*Structural*),



Obrázok 174 Priradenie štrukturálneho tagu pomocou tlačidla

3. kliknite ľavým tlačidlom myši na prázdne okno v stĺpci Štruktúra (*Structure*) na záložke Rozloženie (*Layout*) a na záložke Štrukturálne tagy (*Structural*) a vyberte požadovanú možnosť.

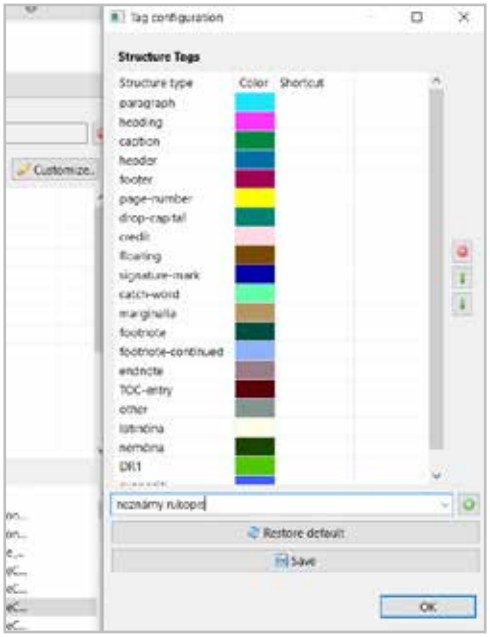
Dvojitým kliknutím na jednotlivý riadok v tejto sekcii sa priblíži požadovaná oblasť v obrázku dokumentu na pracovnej ploche a naopak.



Obrázok 175 Priradenie štrukturálneho tagu pomocou stĺpca Štruktúra (Structure)

7.2.2 Vytvorenie štrukturálneho tagu

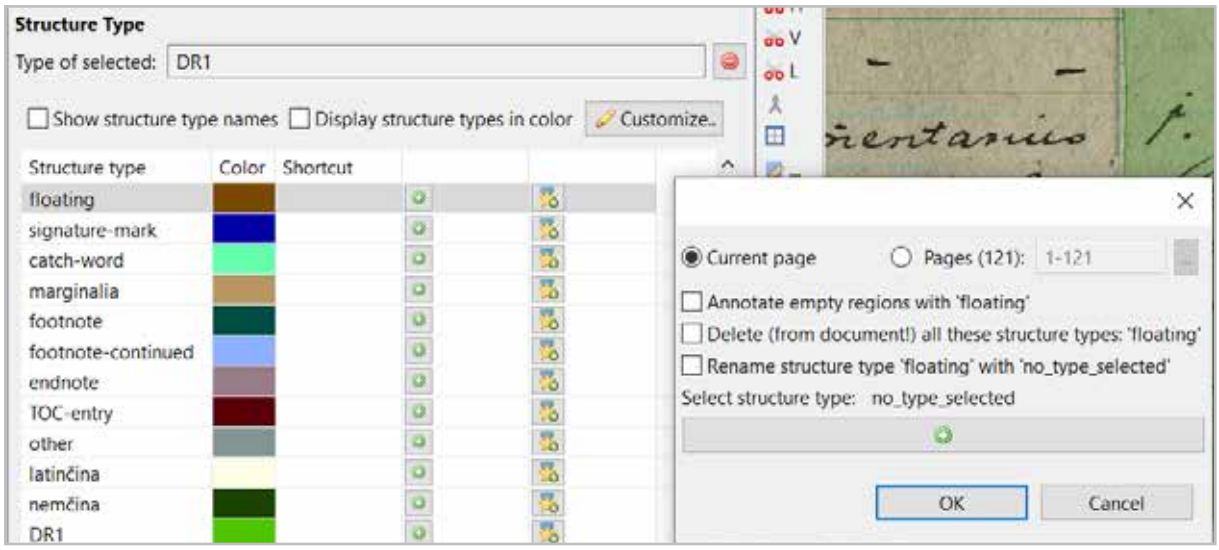
Okrem preddefinovaných štrukturálnych tagov si môžete vytvoriť vlastné tagy kliknutím na tlačidlo Prispôsobiť (*Customize*). Po otvorení okna Konfigurácia tagu (*Tag configuration*) zadajte názov tagu do prázdneho poľa v spodnej časti okna a kliknite na zelené tlačidlo +.



Obrázok 176 Pridanie nového štrukturálneho tagu

Prispôsobiť si môžete aj farebné označenie tagov kliknutím na farebnú časť vedľa štítku a vybrať požadovanú farbu.

Rozšírené možnosti jednotlivých tagov ponúka tlačidlo s hviezdičkou nachádzajúce sa vedľa každého z nich. Voľby sú viazané na zvolený tag: označenie všetkých oblastí prázdneho textu, odstránenie zo všetkých snímok dokumentu, premenovanie.

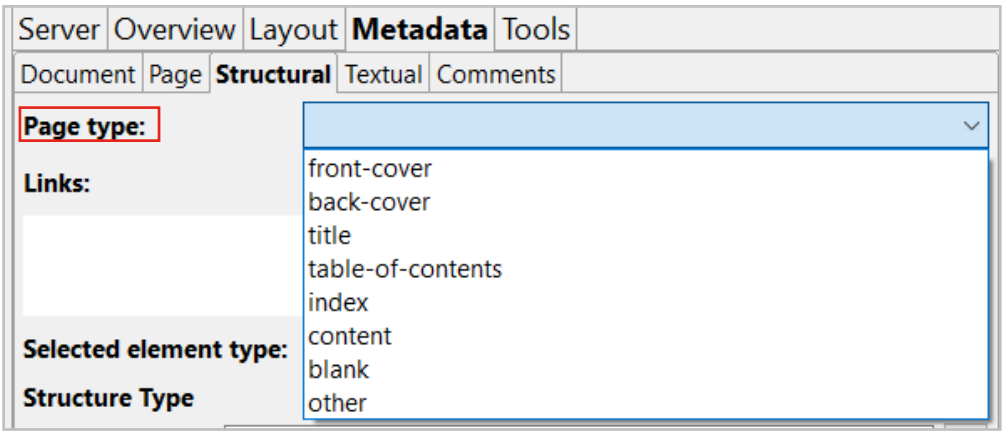


Obrázok 177 Ponuka tlačidla s hviezdičkou

Odstrániť štrukturálny tag môžete prostredníctvom záložky Štrukturálne tagy (*Structural*). Výberom požadovanej oblasti a následne zakliknutím červeného tlačidla na záložke Typ štruktúry (*Structure Type*) alebo na záložke Rozloženie (*Layout*) cez stĺpec Štruktúra (*Structure*) a zaklknutím voľby zmazania vo vyrolovanom okne.

7.2.3 Ďalšie možnosti záložky štrukturálnych tagov

Ku každej strane dokumentu je možné priradiť typ strany výberom z prednastavených možností (predná obálka, zadná obálka, názov, register, obsah, prázdne, iný) v časti **Typ strany** (*Page type*). Kliknite na šípku v prázdnom okne a vyberte požadovaný typ strany vždy pri aktuálnej snímke.



Obrázok 178 Voľby v časti Typ stránky (Page type)

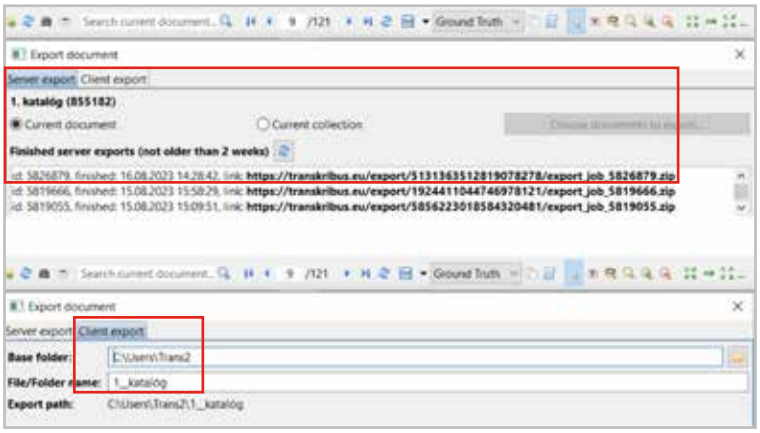
Štrukturálne tagy môžete prepojiť prostredníctvom tlačidla **Prepojenia** (*Links*), napríklad prepojenie riadku s poznámkou pod čiarou a pod.

7.3 Export výstupov

So svojimi obrázkami a prepismi môžete pracovať aj mimo platformy. Slúži na to export výstupov. Rôzne funkcie vám umožnia prispôsobiť výstup podľa formátu súboru a možností, ktoré uprednostňujete. Môžete si vybrať export na úrovni strán, obrázkov, textu alebo štrukturálnych prvkov.

Funkcie pre export nájdete kliknutím na ikonu Exportovať dokument (*Export document*), ktorá sa nachádza na hlavnom paneli. V zobrazenom okne najprv vyberte z ponuky záložiek, ktoré rozhodujú o spôsobe uloženia, pričom jeho priebeh môžete sledovať v okne *Jobs* na záložke *Server*. Možnosti:

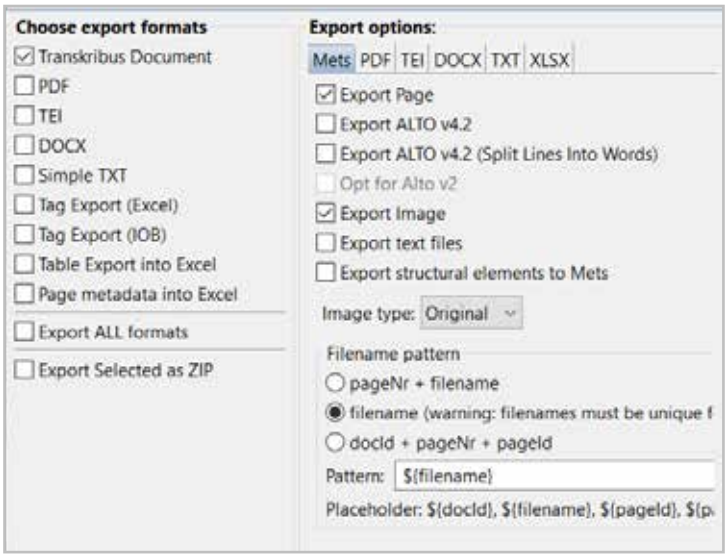
1. export zo servera (*Server export*) pre výstup zo servera platformy s odkazom na stiahnutie,
2. export z klienta (*Client export*) pre výstup do počítača na zvolené miesto.



Obrázok 179 Okná so záložkami Export zo servera (Server export) a Export z klienta (Client export)

7.3.1 Voľba formátu

Následne sa rozhodnite pre vhodný formát. Výber je dvojkrokový, najprv vyberte základný formát v stĺpci Vybrať formát exportu (*Choose export formats*) a potom zvoľte rozšírené možnosti v druhom stĺpci Možnosti exportu (*Export options*).

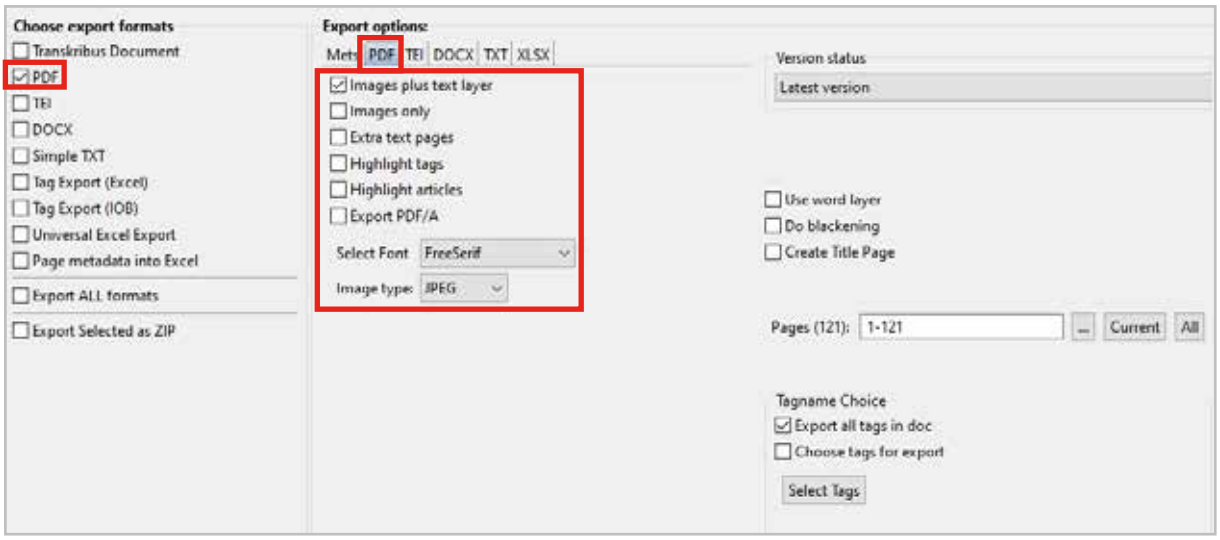


Obrázok 180 Ponuka možností pre formáty exportu

Formát Transkribus Document
Pri zakliknutí tejto možnosti sa vytvorí súbor METS, ktorý obsahuje odkazy na rôzne súbory v závislosti od toho, ktorú možnosť si vyberiete (Page, ALTO, obrázky, text, tagy). Obsahuje všetky základné informácie o súbore.

Formát PDF
Z ponuky si môžete vybrať medzi viacerými možnosťami:

1. Obrázky plus textová vrstva (*Images plus text layer*) – zobrazí sa prepísaný text a obrázok dokumentu.
2. Iba obrázky (*Images only*) – prepísaný text nebude viditeľný.
3. Dodatočné textové stránky (*Extra text pages*) – prepísaný text sa pridá ako ďalšia strana po každom obrázku.
4. Zvýrazniť tagy alebo články (*Highlight tags/Highlight articles*) – zvýraznené údaje sa zobrazia vo farbách používaných na platforme a na konci dokumentu sa vygeneruje legenda vysvetľujúca význam rôznych farieb.
5. Export v PDF/A (*Export PDF/A*) – na dlhodobé uchovanie.



Obrázok 181 Ponuka pre formát PDF

Formát TEI
Táto možnosť je určená pre používateľov z komunity konzorcia Text Encoding Initiative (TEI).

Formát textových súborov
Na výber sú k dispozícii súbory programu Word (docx), kde si môžete vybrať možnosti týkajúce sa zlomov riadkov, skratiek a ďalších alebo sa môžete rozhodnúť pre jednoduchý súbor txt. V tomto prípade môžete vytvoriť súbory triedené podľa tagov, od prvého po posledný, pričom tieto súbory môžu byť pomenované podľa jedného alebo viacerých atribútov tagu.

Export podľa tagov
Pre export priradených tagov existujú tri možnosti:

1. súbor Excel – vytvorí sa súbor s jednotlivými záložkami pre každú kategóriu tagov a jednou záložkou s prehľadom všetkých tagov,
2. súbor PDF – zvýraznia sa tagy v exportovanom súbore,
3. súbor docx – tagy sú viditeľné v exportovanom súbore.

7.3.2 Ďalšie možnosti

Stav verzie (*Version status*) – táto možnosť umožňuje exportovať jednotlivé (predchádzajúce) verzie dokumentu.

Slovná vrstva (*Word Layer*) – exportuje sa text zo segmentácie slovnej vrstvy. Funguje len vtedy, ak ste počas rozpoznávania textu zvolili možnosť Pridať odhadované súradnice slova (*Add estimated word coordinates*).

Vykonať začiernenie (*Do blackening*) – táto možnosť funguje len pre súbory Word, PDF a METS. Začiernené citlivé časti prepisu zostanú skryté aj v exportovaných súboroch.

Vytvoriť titulnú stranu (*Create title page*) – titulná strana sa vygeneruje z informácií pridaných na záložkách Dokument (*Document*) a Metadáta (*Metadata*). Môžete sem vložiť informácie o názve, autorovi, jazyku a dátume dokumentu. Môžete tiež vytvoriť redakčné vyhlásenie (*Editorial declaration*) popisujúce postupy pri transkripcii dokumentu.

Posledným krokom je výber počtu strán, ktoré chcete exportovať. Môžete exportovať všetky strany, vybrané strany, rozsah alebo len aktuálnu stranu.

8 Základy automatickej transkripcie na platforme Transkribus Lite

Platforma Transkribus Lite (v súčasnosti prezentovaná pod názvom Transkribus web app) je vo svojej podstate bezplatná webová verzia softvéru na automatickú transkripciu rukopisných alebo tlačených dokumentov Transkribus Expert Client. Mnohé z používateľmi obľúbených funkcií v prostredí Transkribus expert klient je možné s drobnými variáciami a istými obmedzeniami nájsť a použiť aj v Transkribus Lite. Aj na webovej platforme Transkribus Lite je teda možné po vytvorení si konta vložiť do systému digitálnu kópiu jedného historického dokumentu alebo niekoľkých rôznych dokumentov), a to v podobe digitálnych snímok (skenov) alebo PDF dokumentu. Na automatický prepis týchto digitálnych kópií je možné použiť sprístupnené špecifické nástroje, tzv. modely rozpoznávania rukopisného textu (*HTR Models*). Na účel automatickej transkripcie môže byť použitý aj užívateľom vytvorený vlastný, teda špecifický model rozpoznávania rukopisného/tlačeného textu. Okrem získania a uplatnenia modelu na automatický prepis vloženého historického dokumentu je na platforme Transkribus Lite samozrejmosťou aj vyhľadávanie vybraného slova či skupiny slov (slovných spojení) vo vloženom digitalizovanom dokumente (dokumentoch) a jeho automatickej transkripcii.

8.1 Webové umiestnenie a výzor stránky platformy Transkribus Lite

Platforma Transkribus Lite je dostupná na webovej stránke https://app.transkribus.eu/

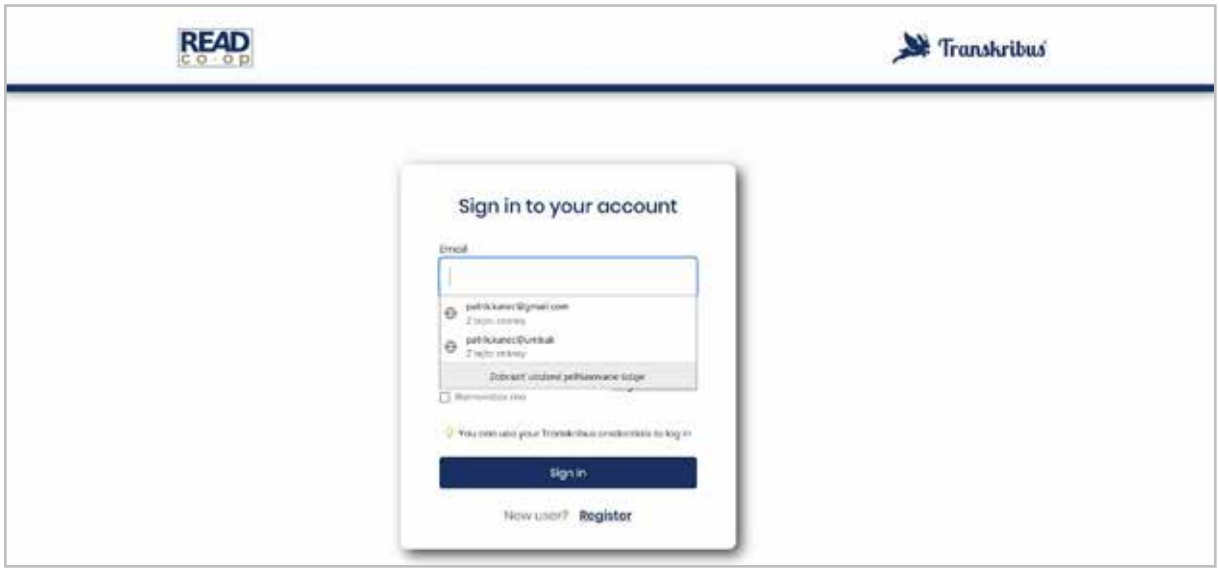
Obrázok 182 Úvodná stránka Transkribus Lite verzia 2.3.0.8 používaná do 30. augusta 2023

30. augusta 2023 bola webová stránka platformy Transkribus Lite prepracovaná, došlo nielen k vizuálnej zmene stránky (zmena tzv. front-endu), ale aj k zmene a doplneniu obsahu a niektorých funkcií. Postupy, ktoré približujeme v nasledujúcich riadkoch, sú vypracované na podklade najnovšej podoby webstránky Transkribus Lite.

*Poznámka: V priebehu prvej polovice septembra 2023 sme pracovali s verziami 3.0.0.14 až 3.0.0.18. Platforma Transkribus Lite je teda stále v procese vylepšovania. V spodnej časti úvodnej obrazovky platformy Transkribus Lite bolo do 30. augusta možné otvoriť si užívateľskú príručku/interaktívny manuál (*Getting started with Transkribus Lite*). V novej verzii platformy Transkribus Lite (verzia 3.0.1.22) je táto možnosť prístupná cez záložku* Help *na spodnej lište stránky. Parciálne vysvetlivky sú dostupné v jednotlivých krokoch príprav dokumentov na seg-*

mentáciu a automatickú transkripciu – väčšinou v podobe znaku „i“ (= informácia), prípadne sa potiahnutím myši nad ikonou zobrazí vyskakovacie okno s vysvetlením.

Úvodná obrazovka platformy Transkribus Lite ponúka užívateľovi vytvoriť si **bezplatné konto**. Stačí uviesť e-mailovú adresu a vytvoriť si heslo. Novovytvorené konto je potrebné si aktivovať v automaticky zasielanom e-maili, ktorý sa odosiela na užívateľom uvedenú e-mailovú adresu. Používateľ, ktorý má vytvorené konto na platforme Transkribus Expert Client, sa môže automaticky prihlásiť aj do prostredia Transkribus Lite.



Obrázok 183 Stránka platformy Transkribus Lite po vytvorení používateľského konta

Po zvolení si užívateľského konta (užívateľ môže mať viac kont, ako je evidentné aj z nášho príkladu) a zadaní hesla (aj s pomocou automatického vkladania hesla webovým prehliadačom) sa užívateľ dostane do **pracovného prostredia** Transkribus Lite.

Obrázok 184 Základné (úvodné) prostredie po prihlásení

Na obrázku vyššie už konto platformy Transkribus Lite obsahuje aj jednu užívateľom vloženú zbierku digitalizovaného dokumentu s názvom *DA BBD – Prothocollum Ecclesiae Radvanensis*.

Vloženie vlastného dokumentu do zbierky je vysvetlené v kapitole *8.4 Vytvorenie zbierky dokumentov a nahratie dokumentov*.

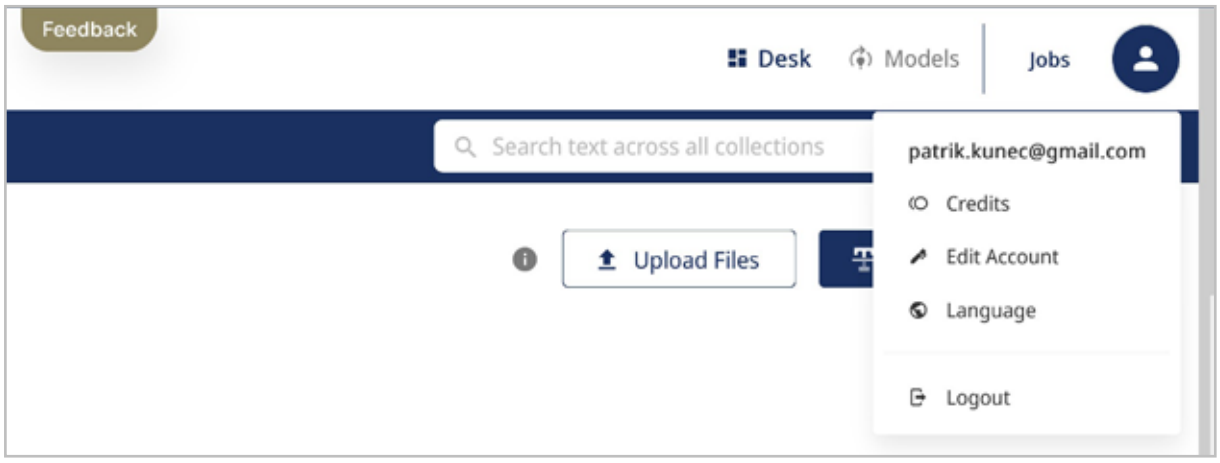
8.2 Registrácia a cena za prístup k službám Transkribus Lite

Na platforme Transkribus Lite je vytvorenie konta bezplatné. Bezplatné konto umožňuje na platformu nahrať digitálnu podobu historického dokumentu, vytvoriť jeho textové rozloženie (segmentácia) pre potreby manuálneho alebo automatického prepisu časti alebo aj celého dokumentu. Bezplatná verzia umožňuje použiť už vytvorené a sprístupnené modely automatickej transkripcie, ako aj vytvorenie si vlastného modelu automatického rozpoznávania textu. Za použitie automatického transkripcie modelom HTR používateľ platí tzv. kreditmi, ktoré získava zdarma pri registrácii (v čase tvorby tejto metodickej príručky to bolo 500 kreditov). Kreditmi sa platí za automatickú segmentáciu a transkripciu každej strany dokumentu. Stav kreditov sa priebežne aktualizuje a zobrazuje sa v prehľade užívateľského konta (*Credits*).

8.3 Základná pracovná plocha

Po prihlásení sa do prostredia Transkribus Lite sa zobrazí hlavná plocha pracovného prostredia. Pracovná plocha má viac-menej jednoduchú štruktúru, ktorá umožňuje aj intuitívne oboznámenie sa s funkciami platformy.

V pravej hornej časti sa nachádza krátka lišta s možnosťou zobrazenia a zmien pracovného prostredia, ktorá obsahuje tri textové tlačidlá. Tie umožňujú zobraziť základnú plochu so zbierkou/zbierkami dokumentov (*Desk*), bezplatne sprístupnené modely automatickej transkripcie (*Models*) a prehľad práve prebiehajúcich úloh (*Jobs*). Za posledným textovým tlačidlom sa nachádza ikona používateľa konta, ktorá skrýva možnosť zobrazenia kreditov, editovania konta, výber jazyka a odhlásenie sa z konta.



Obrázok 185 Lišta s textovými tlačidlami a kontom používateľa

Na lište s modrým podfarbením sa nachádzajú tri možnosti zobrazenia pracovného prostredia – základné pracovné prostredie (*Home*), náhľad na zbierky digitalizovaných dokumentov (*Collections*) a napokon možnosť vytvárania špecifických značiek a poznámok (*Tags*). V pravej časti lišty sú dve vyhľadávacie okná – na vyhľadávanie slova alebo slov v zbierke všetkých dokumentov (*Search text across all collections*) a všeobecný textový vyhľadávač (*Global Text*

Search), ktorý ponúka možnosť vyhľadať zvolené slovo buď vo vybranej zbierke alebo v konkrétnom dokumente.

V pravej hornej časti pracovnej plochy sa nachádzajú ešte dve tlačidlá, ktoré slúžia na:

- nahratie súborov digitalizovaných dokumentov (*Upload Files*),
- rýchle rozpoznanie textu s jeho automatickou transkripciou vo vybranom modeli (*Quick Text Recognition*).

V ľavej časti pracovnej plochy sú pod uvítacou formulou vyobrazené náhľady na súbory digitalizovaných dokumentov, s ktorými užívateľ naposledy pracoval. V dolnej časti základnej pracovnej plochy sa ešte zobrazujú jednotlivé zbierky v konte užívateľa.

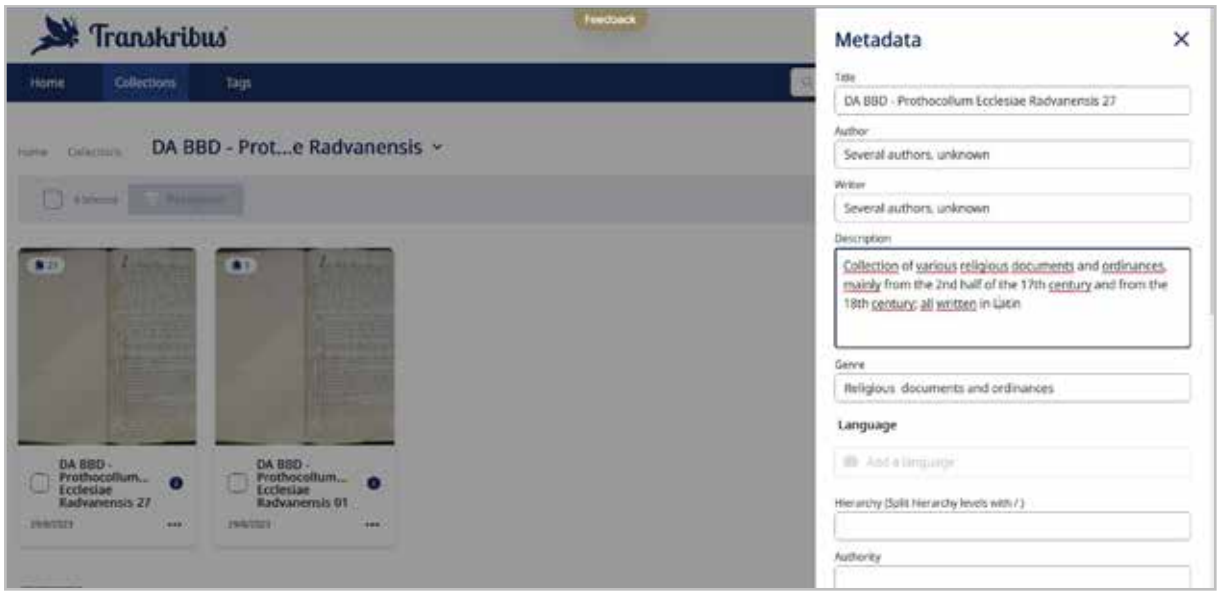
Obrázok 186 Hlavná pracovná plocha po prihlásení s náhľadom na dokument s prameňom, s ktorým užívateľ naposledy pracoval

8.4 Vytvorenie zbierky a nahratie dokumentov

Podobne ako na platforme Transkribus Expert Client, aj vo verzii Transkribus Lite si používateľ môže vložiť do pracovného prostredia viaceré súbory digitalizovaných dokumentov, s ktorými môže na platforme ďalej pracovať. Podporované sú digitalizáty vo formátoch JPEG, PNG a PDF v čo najvyššom rozlíšení. **Odporúčané rozlíšenie pri obrázkoch je aspoň 300 DPI.**

Súbory s digitalizovanými dokumentmi sa nahrávajú z pamäte pracovného zariadenia (počítača) kliknutím na tlačidlo *Upload Files*. Systém automaticky ponúkne možnosť vyhľadať priečinok s digitalizátmi v pracovnom zariadení a jednoducho ho nahrať na platformu. Následne je potrebné súbor digitalizovaných dokumentov s prameňom pomenovať.

Ku každej zbierke digitalizovaných dokumentov je potrebné vytvoriť krátky popis, tzv. metadáta. Formulár obsahuje riadky s možnosťou zadať názov prameňa, autora a pisára (viacerých pisárov), krátky opis obsahu, datovanie, žáner a jazyk.



Obrázok 187 Formulár pre vytvorenie metadát k zbierke digitalizovaných prameňov

Obsah jednotlivých zbierok digitalizovaných dokumentov je možné otvoriť kliknutím na ich náhľadové okno, či už na základnej pracovnej ploche alebo otvorením záložky *Collections*. Systém po otvorení zvolenej zbierky digitalizátov ponúkne náhľad na všetky digitalizované jednotky v prehľadnom zobrazení.

Obrázok 188 Prehľadné zobrazenie digitalizovaných prameňov v jednom priečinku (v parciálnej zbierke)

K otvorenej zbierke vybraného digitalizovaného dokumentu je možné vložiť ďalšie snímky, a to:

- pretiahnutím konkrétnej digitalizovanej strany z pracovného zariadenia priamo do okna s jednotlivými digitalizátmi, alebo
- vyhľadaním snímky pomocou prehliadania súboru s digitalizátmi v pracovnom zariadení (*Browse/Browsing*) a jej vložením pomocou tlačidla Nahrať (*Upload*).

Jednotlivé snímky sa dajú aj odstrániť. Túto možnosť ponúka menu s tromi bodkami v náhľade každej snímky. Okrem toho je v tomto menu možné vytvoriť metadáta pre každú

jednotlivú snímku v zbierke digitalizovaného dokumentu (túto funkciu už bližšie nepredstavujeme).

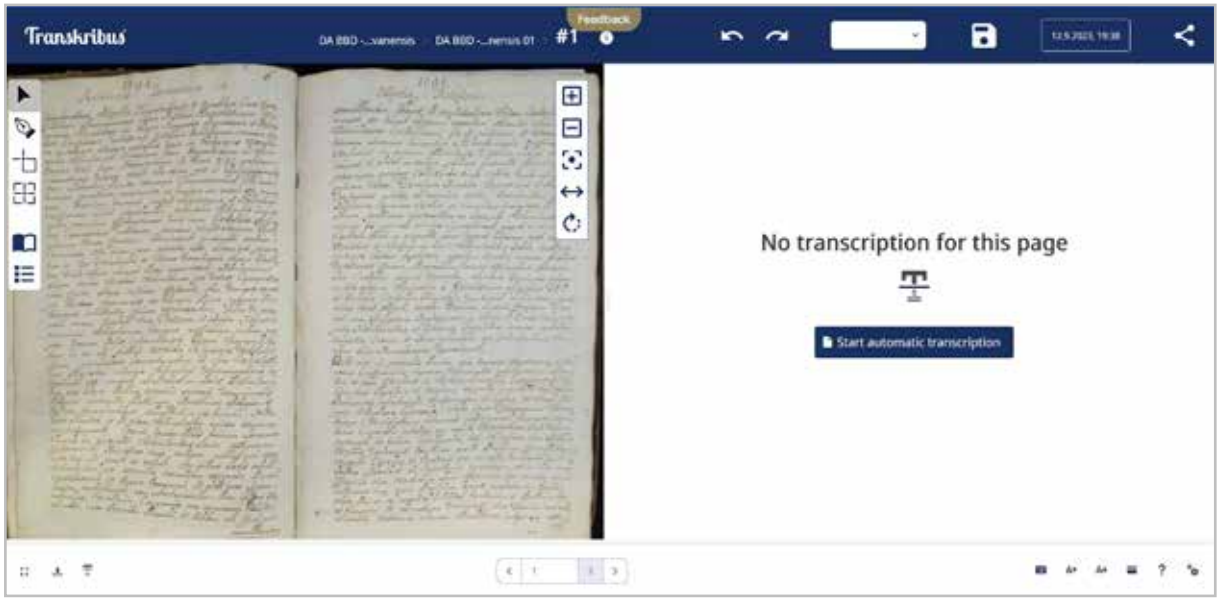
Práca s nahrávaním a popisovaním zbierok digitalizovaných dokumentov je pomerne jednoduchá a na jej pochopenie stačia základné znalosti práce so súbormi, ich sťahovaním a popisovaním.

8.5 Segmentácia a automatická transkripcia vybraného dokumentu v zbierke

Základnou funkciou platforiem Transkribus Expert Client a Transkribus Lite je možnosť automatickej transkripcie do systému vložených rukopisných alebo tlačených digitalizovaných dokumentov. Procesu automatickej transkripcie predchádza dôležitý krok, ktorým je segmentácia textu prameňa (tiež analýza rozloženia textu) v každej jednotlivej snímke do textových rámcov, riadkových rámcov alebo tabuliek. Až po segmentácii textu prameňa je možné prejsť k automatickej transkripcii na podklade použitia jedného z voľne dostupných modelov rozpoznávania rukopisných/tlačených textov (*HTR Models*).

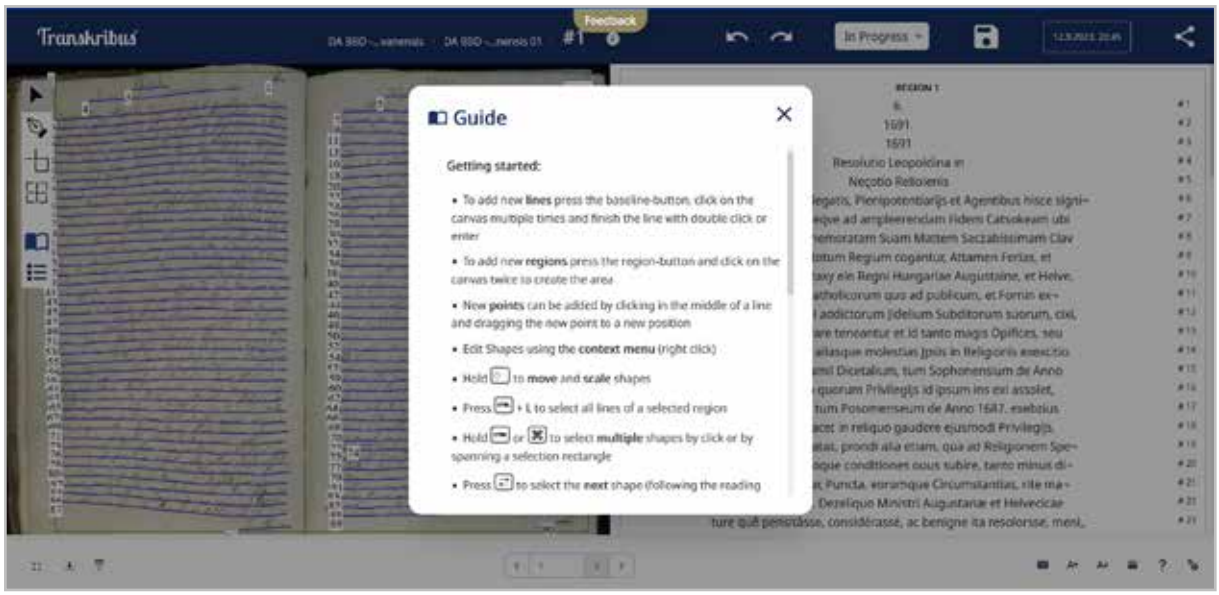
Po kliknutí na náhľad jednotlivej snímky v zbierke sa zobrazí úvodné okno s tlačidlom na spustenie automatickej segmentácie a transkripcie (*Start automatic transcription*).

Z nahraného prameňa s názvom *DA BBD – Prothocollum Ecclesiae Radvanensis* sme vybrali tretiu dvojstranu, na ktorej demonštrujeme proces segmentácie a automatickej transkripcie.



Obrázok 189 Vybraný digitalizovaný prameň pred spustením segmentácie a automatickej transkripcie

Na obrázku vidieť, že ľavá časť pracovnej plochy ponúka obraz digitalizovaného prameňa (v tomto prípade dvojstránka), ktorý je možné segmentovať podobne ako na platforme Transkribus expert klient, ale s použitím iných ikon (ich funkcia sa zobrazí presunutím myši nad ikonu). Na ľavej strane je môžete zvoliť možnosť Pridať riadok (*Add Line*), pridať textový rámec (*Add Region*), pridať v naskenovanej snímke tabuľku (*Add Table*). Pod týmito pracovnými ikonami sa nachádza ikona sprievodcu (*Guide*) a ikona s príkazom na zobrazenie vykonanej segmentácie v prehľadnej tabuľke Rozloženie (*Layout*). Na pravej strane zobrazovacieho okna s digitalizátom prameňa je ďalších päť ikon, ktoré umožňujú naskenovaný prameň zväčšiť (*Zoom in*), zmenšiť (*Zoom out*), centrovať (*Center*), natiahnuť do všetkých strán zobrazovacieho okna (*Fit to width*) a otáčať (*Rotate*).

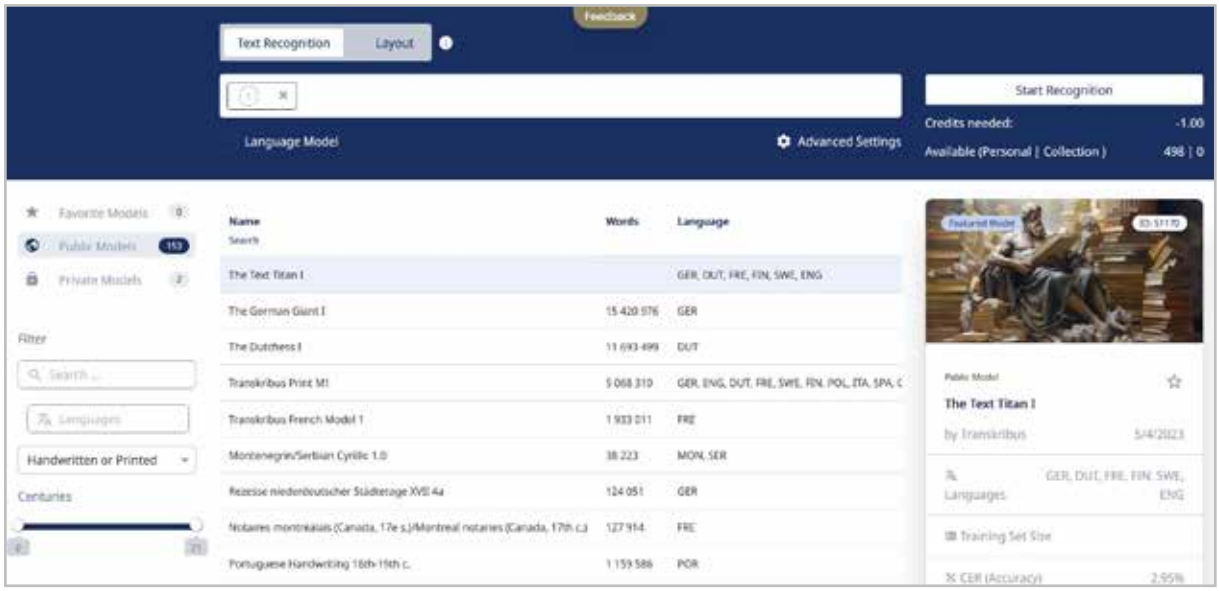


Obrázok 190 a 191 Detailnejšie vysvetlenie jednotlivých možností segmentácie poskytuje sprievodca (Guide)

Systém segmentácie ponúka všetky utility ako v prostredí Transkribus expert klient s tým rozdielom, že sa realizujú pomocou kombinovaných klávesových skratiek.

Môžete realizovať vlastnú, manuálne upravenú segmentáciu textu, alebo spustiť automatickú segmentáciu vrátane automatickej transkripcie. Túto možnosť vyberte jednoduchým kliknutím na tlačidlo Spustiť automatickú transkripciu (*Start automatic transcription*) na pravej strane pracovnej plochy.

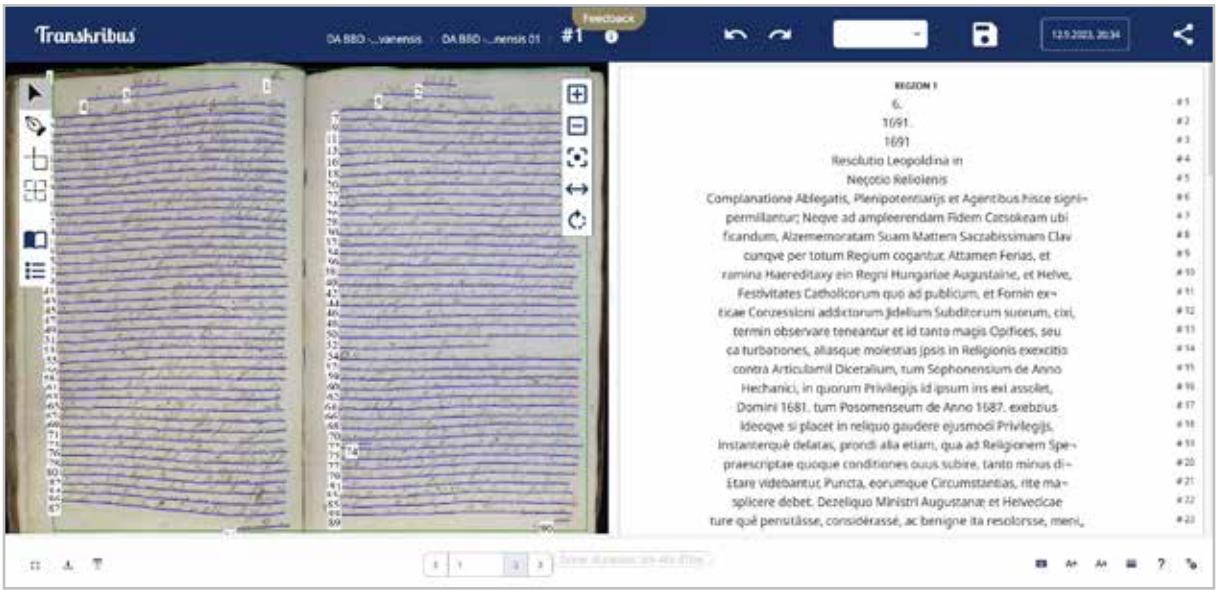
Po výbere tejto možnosti systém Transkribus Lite v novom okne ponúkne výber z bezplatne sprístupnených modelov vytvorených pre automatický prepis prameňov rôzneho obsahu, jazykov a tiež z rôznych historických epoch (v čase tvorby tohto textu bolo k dispozícii 153 modelov). Pre potreby demonštrácie ďalších pracovných postupov sme vybrali prvý ponúkaný model s názvom *The Text Titan I*, ktorý má pomerne veľkú mieru presnosti prepisu (indikátor CER je 2,95 %; bližšie o hodnotení kvality modelu v kapitole *5.3.1 Hodnotenie úspešnosti modelu*), keďže v čase tvorby tohto textu sa v ponuke modelov nenachádzal žiadny, ktorý by bol vytvorený na prepis latinsky písaných dokumentov z obdobia 18. storočia ako vzorový prameň.



Obrázok 192 Ponuka modelov automatickej transkripcie

Po výbere voľby Rozpoznanie textu (*Start Recognition*) sa začne vykonávať úloha, ktorá je spoplatnená jedným kreditom. Stav konta s kreditmi sa zreteľne zobrazuje pod tlačidlom na vykonanie rozpoznania textu). Trvanie zadanej úlohy je možné sledovať na záložke Úlohy (*Jobs*) na hlavnej pracovnej ploche užívateľského konta.

Po ukončení zadanej úlohy, t. j. po vykonaní automatickej segmentácie a prepisu digitalizovaného textu prameňa, sa zobrazí okno so segmentovaným textom prameňa (vľavo) a s jeho automatickým prepisom na základe vybraného modelu (vpravo).



Obrázok 193 Výsledok segmentácie a transkripcie na ilustračnom príklade prameňa

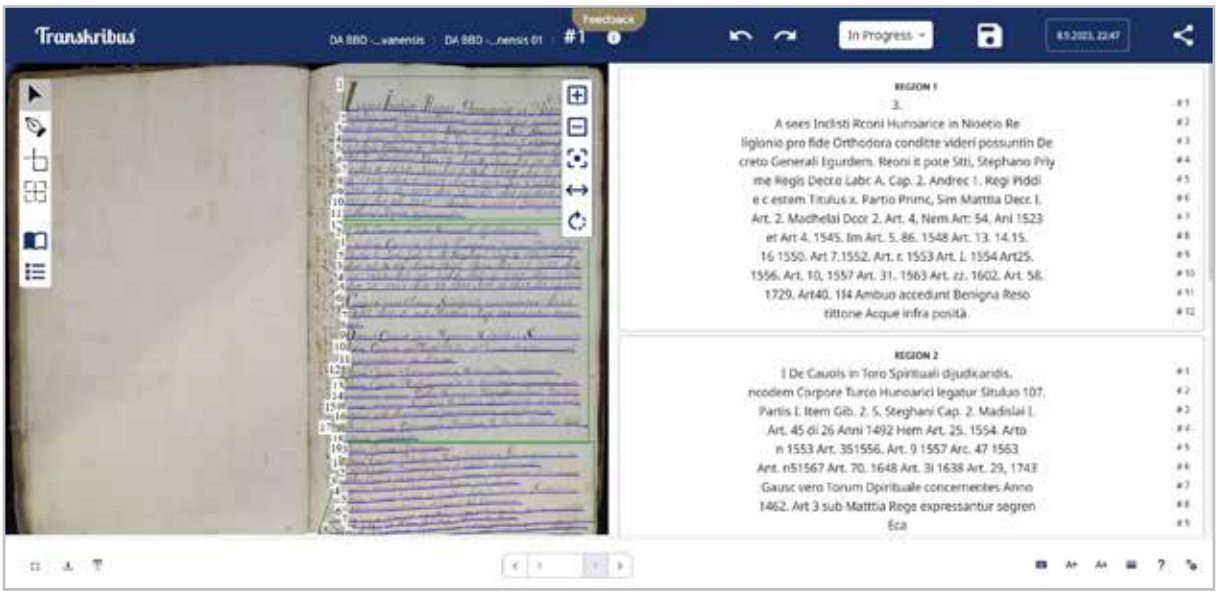
Hoci je na prvý pohľad zrejmé, že automatický prepis na základe zvoleného modelu neurobil veľmi presnú transkripciu latinského textu z polovice 18. storočia, približne 75 – 85 % textu je prepísaného správne. Používateľ si môže v takto segmentovanom a automaticky prepísanom dokumente robiť vlastné úpravy/opravy (či už segmentačné alebo transkripčné) a získať tak čo najpresnejší prepis dokumentu. V upravenom prepise (prípadne vo viacerých prepisoch) je napríklad možné vyhľadávať vybrané slovo alebo slová, a to nielen v jednom súbore digitalizo-

vaného prameňa, ale vo všetkých zbierkach dokumentov. Platforma Transkribus Lite tak ponúka nielen možnosť automatického prepisu, ale aj pridané funkcionality, ktoré uľahčujú ďalšiu prácu s digitalizovanými historickými prameňmi.

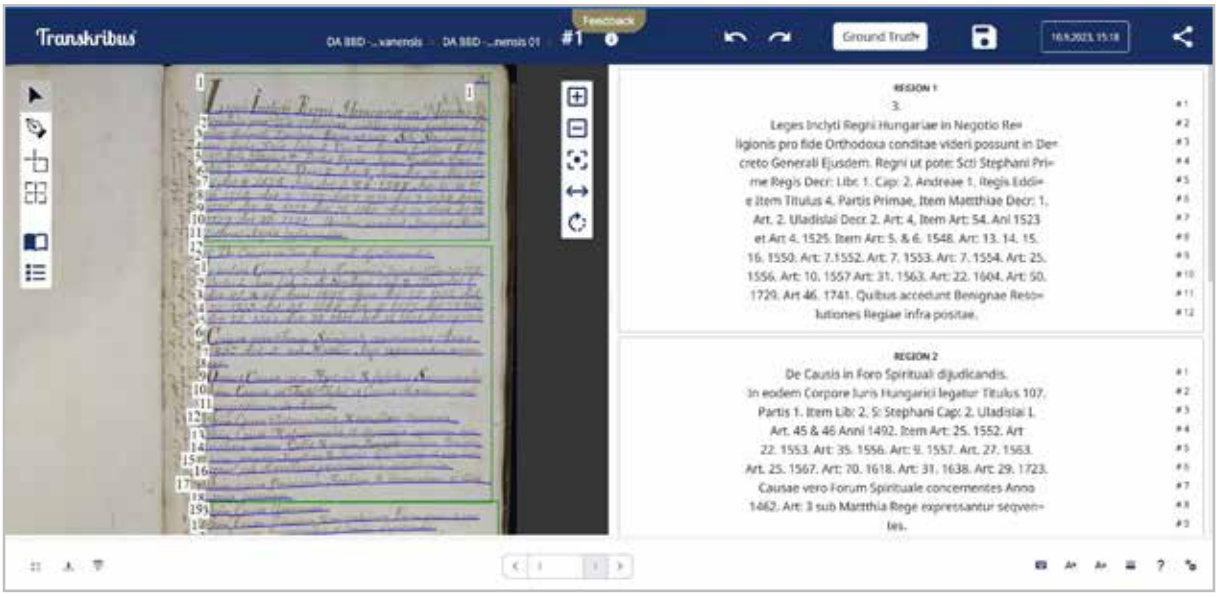
8.6 Tvorba vlastného modelu automatickej transkripcie

Platforma Transkribus Lite umožňuje vytrénovať si na podklade digitalizovaného dokumentu vlastný model automatickej transkripcie. Na jeho vytvorenie sa odporúča prepísať aspoň 20 strán z digitalizovaného obsahu, aby sa softvér na vytváranie modelov automatického prepisu "naučil" čítať konkrétny typ písma digitalizovaného prameňa. Na účely demonštrácie postupu tvorby modelu sme vybrali jeden z digitalizovaných prameňov v osobnom konte, ale obmedzili sme prípravu cvičného súboru len na prepis dvoch snímok (troch strán textu). Je evidentné, že na základe malého počtu prepísaných strán bude mať vzorový model veľmi vysokú mieru chybovosti znakov (CER).

Pre potreby vytrénovania modelu boli vybrané prvé dve snímky z prameňa *Prothocollum Ecclesiae Radvanensis*, ktorý obsahuje prepisy rôznych cirkevných nariadení a predpisov z 17. a 18. storočia. Rukopisný dokument je písaný v latinskom jazyku, typ písma je možné označiť za typickú barokovú podobu humanistickej kurzívy. Naskenovaný text prvej strany prameňa bol najprv automaticky transkribovaný modelom *The Text Titan I*, ktorý sa ukázal ako celkom presný. Následne bolo potrebné automatický prepis opraviť do správnej textovej podoby. Takto opravenú verziu sme označili ako *Ground Truth*.

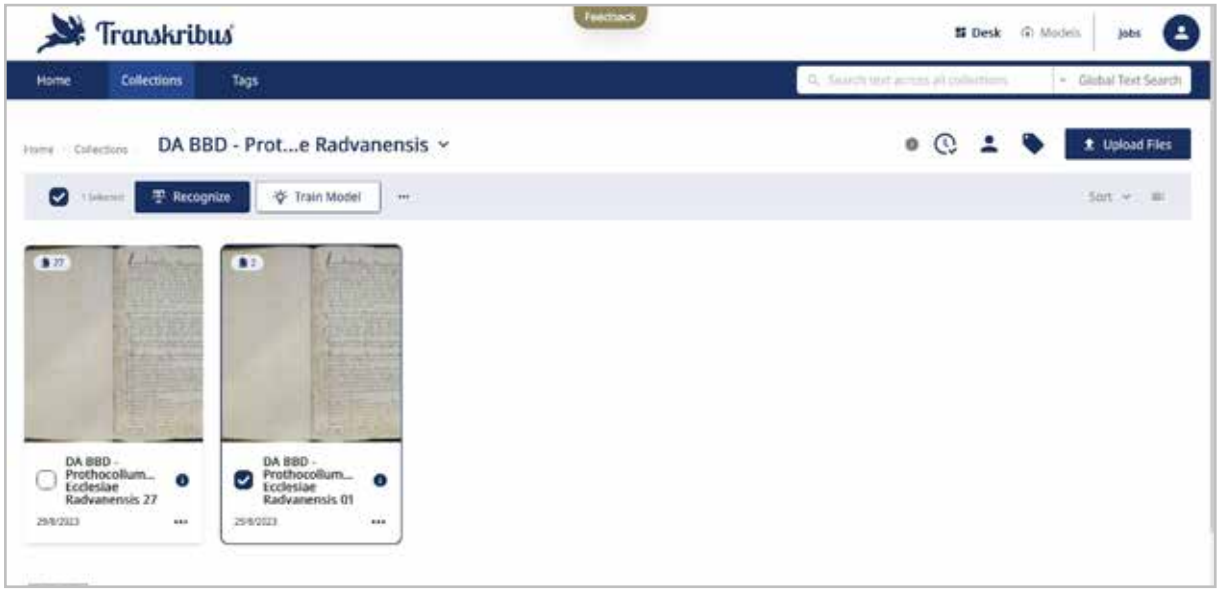


Obrázok 194 Text prvej strany vybraného prameňa po automatickej transkripcii vybraným modelom vo fáze In Progress



Obrázok 195 Text prvej strany vybraného prameňa po textovej úprave a uložení prepisu ako Ground Truth verzie

Rovnakým spôsobom bola realizovaná automatická transkripcia druhej snímky vybraného prameňa, ktorá obsahovala dve strany textu a viac ako 500 slov. Oba tieto automaticky transkribované a textovo upravené pramene boli uložené do nového súboru s názvom *DA BBD – Prothocollum Ecclesiae Radvanensis 01* v zbierke prameňov.

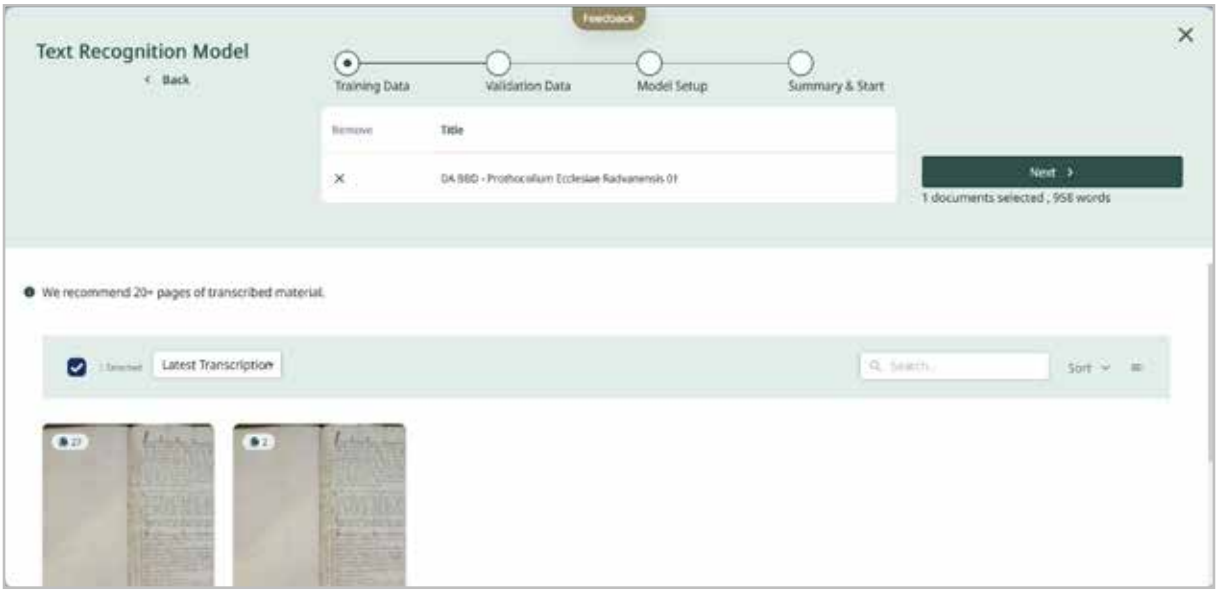


Obrázok 196 Pracovná plocha po výbere zvoleného výberu z prameňa

Po výbere súboru dvoch segmentovaných a prepísaných snímok je možné pristúpiť k trénovaniu modelu použitím tlačidla Trénovať model (*Train Model*).

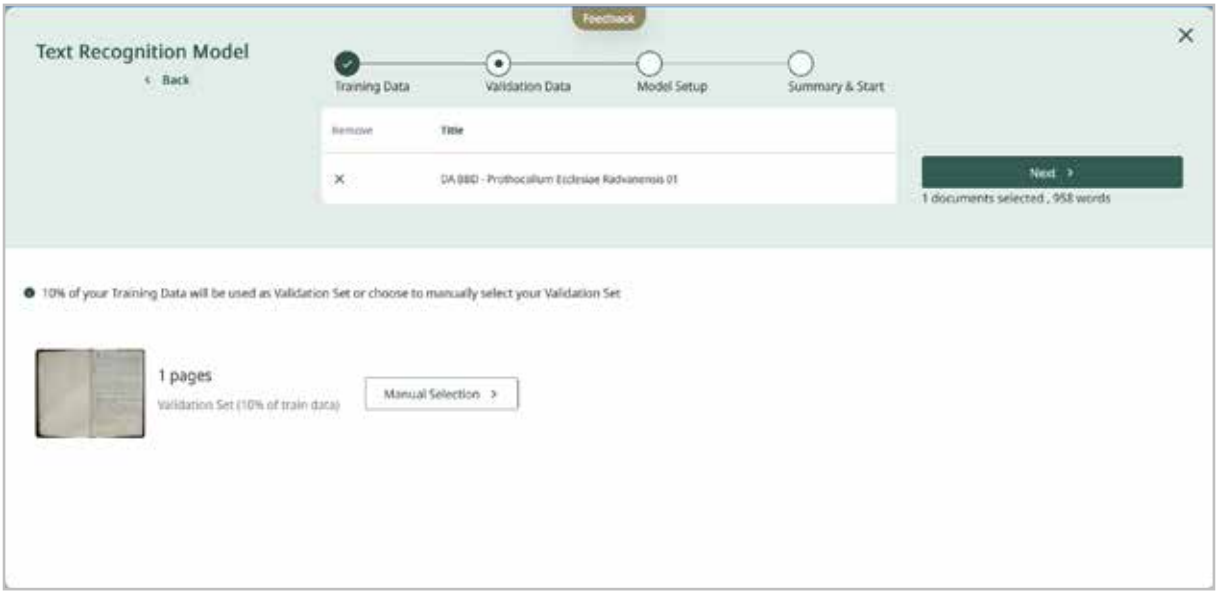
Postup vytvárania nového modelu pozostáva zo štyroch krokov, ktoré sa zobrazujú v štyroch postupne otváraných oknách:.

1. zobrazí sa informácia, že na vytvorenie modelu bude slúžiť vybraný cvičný súbor dát (*Training Data*) s rozsahom dve snímky a s celkovým počtom 958 slov.



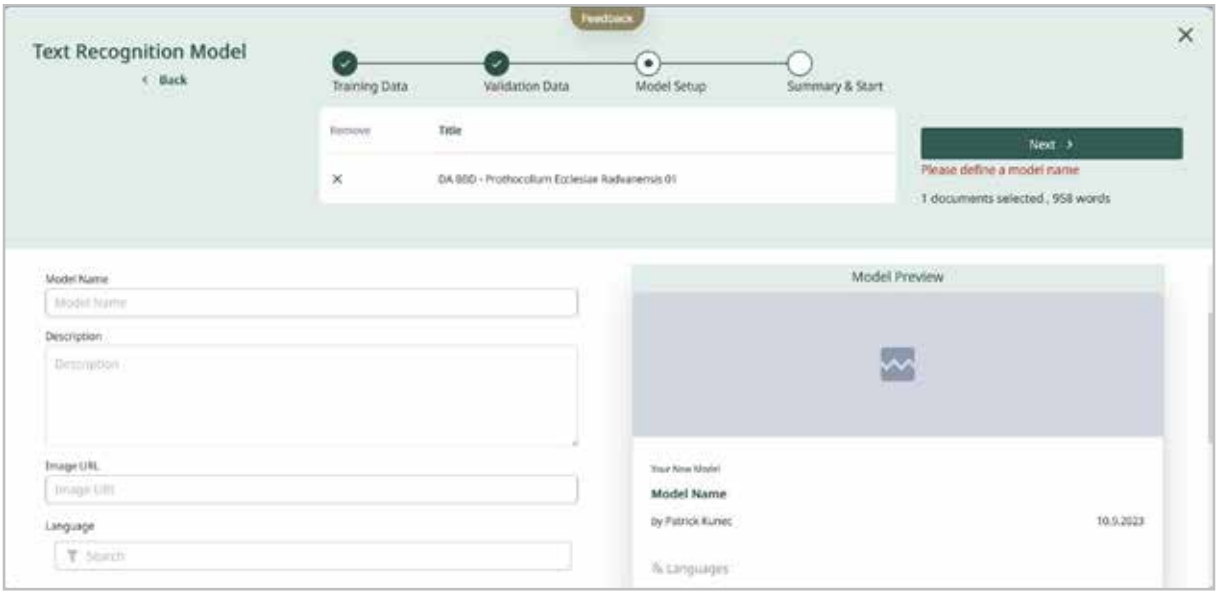
Obrázok 197 Náhľad na prvý krok pri tvorbe vlastného modelu

2. ďalšie okno umožňuje užívateľovi vybrať overovacie dáta (*Validation Data*), teda časť textu z prameňa, na ktorom bude odskúšaná úspešnosť automatického prepisu trénovaného modelu. V prípade vzorového modelu bol ponechaný navrhovaný postup výberu 10 % z cvičného súboru (teda necelých sto slov).

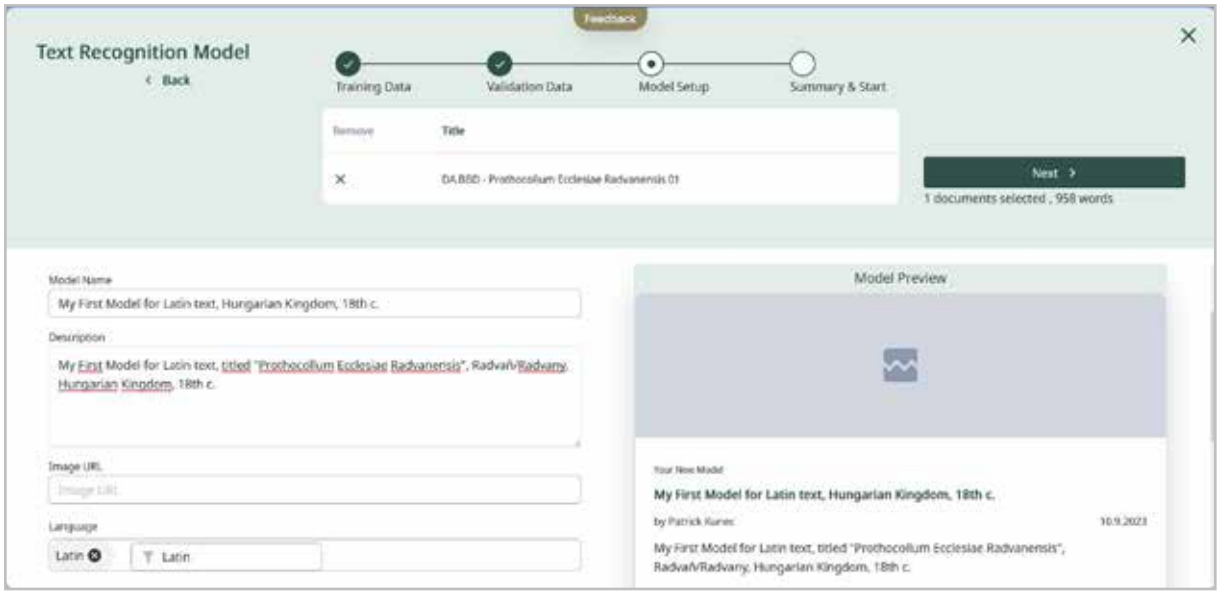


Obrázok 198 Náhľad na druhý krok pri tvorbe vlastného modelu

3. v tejto fáze je možné pomenovať trénovaný model a upraviť niektoré metadáta pomocou funkcie Nastavenia modelu (*Model Setup*).

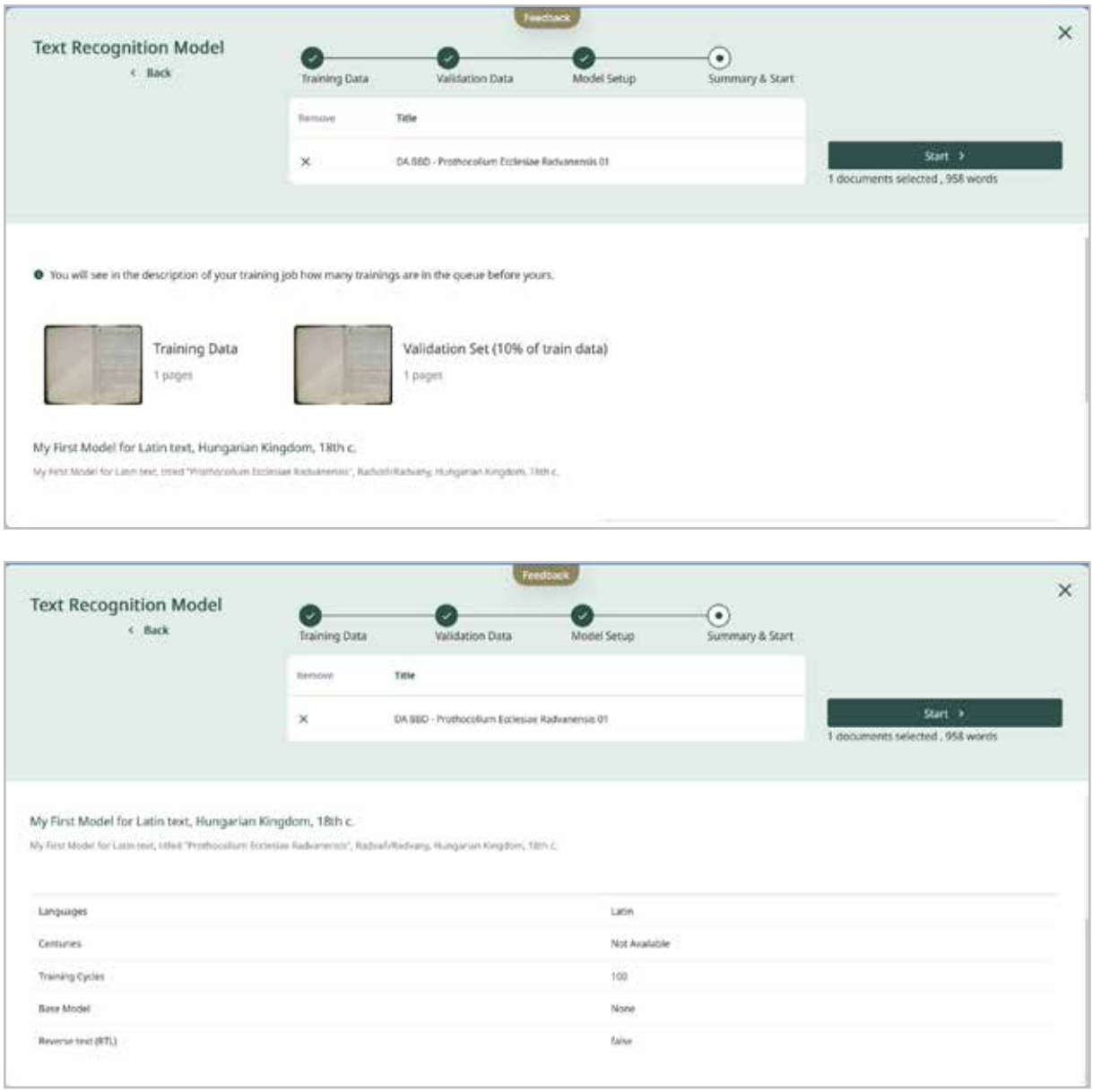


Obrázok 199 Náhľad na tretí krok pri tvorbe vlastného modelu



Obrázok 200 Náhľad na vyplnený formulár v treťom kroku vytvárania vlastného modelu

4. systém ponúkne zhrnutie vložených informácií a možnosť spustenia trénovania vlastného modelu. Vzorový model sme nazvali *My First Model for Latin text, Hungarian Kingdom, 18th c.*. Keďže jazykové prostredie platformy Transkribus Lite je v angličtine, rozhodli sme sa používať anglický jazyk.



Obrázok 201 a 202 Pracovná plocha so štvrtým krokom trénovania modelu

Po kliknutí na tlačidlo Spustiť (*Start*) začne systém trénovať zadefinovaný model automatickej transkripcie. Vzorový model sa trénoval v 100 cykloch a trvalo to niekoľko desiatok minút. Používateľ si môže počet cyklov zvoliť sám, odporúčaných je minimálne 50 cyklov. Riešenie tejto úlohy je možné sledovať na hlavnej pracovnej ploche konta pod záložkou Úlohy (*Jobs*).

Po ukončení trénovania modelu bol vytvorený vzorový model zaradený na novovytvorenú záložku Modely (*Models*), v rámci nej do podzáložky *Dashboard* (horná zelená lište vľavo). Softvér vytvoril pre nový model „identifikačnú kartu“ s vlastným identifikačným číslom (ID modelu). Karta obsahuje základné informácie o modeli vrátane údajov o miere chybovosti v prepise znakov (CER). Vzorový model dosiahol hodnotu CER až 82,2 %, čo je veľmi vysoká miera chybovosti, ale vzhľadom na to, že bol trénovaný len na základe prepisu troch strán z vybraného prameňa, takýto výsledok sa dal očakávať. „Identifikačná karta“ modelu obsahuje v jej spodnej časti aj možnosť graficky zobraziť priebeh vytvárania modelu (*Show Description*).

Obrázok 203 Grafické zobrazenie priebehu trénovania modelu na „identifikačnej karte“ vzorového modelu

Platforma Transkribus Lite umožňuje vytvárať aj ďalšie nové modely, a to použitím tlačidla Trénovať nový model (*Train New Model*) (pozri obrázok č. 202, textové tlačidlo v pravom hornom rohu pracovnej plochy). Rovnako ako v prostredí Transkribus expert klienta aj vo webovom rozhraní Transkribus Lite je možné zvýšiť úspešnosť automatickej transkripcie vo vytvorenom modeli pridaním väčšieho počtu prepísaného textu prameňa do cvičného súboru (odporúčaný počet je aspoň 20 strán).

Vlastné modely automatickej transkripcie, ale aj tie bezplatne sprístupnené, je možné zobraziť a prehliadať si ich „identifikačné karty“ na záložke Galéria (*Gallery*).

Predpokladáme, že verejne prístupných modelov bude na platforme Transkribus Lite časom pribúdať a užívateľ si medzi nimi nájde taký, ktorý mu umožní získať automatický prepis prameňa s minimálnou chybovosťou.

9 Slovník pojmov

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

Canvas (plátno; názov pre menu úprav v Transkribus expert klientovi). Spustenie segmentácie (automatickej analýzy) rozloženia stránky a textu neposkytuje vždy vyhovujúce výsledky. Niekedy sú preto potrebné manuálne korekcie rozloženia. V ponuke Canvas, ktorá sa nachádza spravidla na ľavej strane stránky dokumentu, sa nachádzajú potrebné voľby ako ohraničiť textové rámce (*Text Regions, TR*), pridať riadok (*Lines, L*), pridať základnú čiaru (*Base Line, BL*), pridať slovo (*Word, W*), pridávanie rôznych častí (tabuľky, reklamy, schémy, grafy, grafiky atď.). V ponuke Canvas je tiež možné zmeniť existujúce tvary.

CER (*Character Error Rate*). Miera chybovosti znakov porovnáva pre danú stranu celkový počet znakov (n) vrátane medzier s minimálnym počtom vložení (i), nahradenia (s) a vymazania (d) znakov, ktoré sú potrebné na získanie výsledku Ground Truth. Ide teda o chyby v porovnaní s presným, referenčným textom. Vzorec na výpočet CER: CER = [(i + s + d)/n]*100. Každá malá chyba v prepise je štatisticky plnohodnotná chyba. To znamená, že každá chýbajúca čiarka, „u“ namiesto „v“, dodatočná medzera alebo dokonca veľké písmeno namiesto malého písmena sú zahrnuté v CER ako chyba. Považuje sa za potvrdené a overené konštatovanie, že: a) ak je hodnota chybovosti znakov CER nižšia ako 10 %, čo je 10 a menej chýb na sto znakov, tak výsledok transkripcie je dobrý, čitateľný, a ak je to účelné, je možné ďalšie editovanie výstupu; b) ak je chybovosť znakov CER ≤ 5 %, tak výsledok transkripcie je veľmi dobrý; c) ak je chybovosť znakov CER pod 3 %, potom je možné považovať výsledky transkripcie za výborné a chybovosť znakov CER pod 2,5 % za excelentné.

Cvičný súbor (*Training Set*) pozostáva zo strán, na ktorých sa model trénuje. Na cvičnom súbore sa stroj „učí“, pri každom cykle „prečíta“ rovnakú stranu, pričom chybne prečítané znaky pri každom nasledujúcom cykle vyradí.

DocScan. Open source aplikácia pre Android navrhnutá pre ScanTent. Identifikuje strany dokumentu v živom náhľade a robí snímky v dostatočnej kvalite na transkripciu. V automatickom režime nasníma obrázok po otočení stránky. Umožňuje rýchlo snímať knihy alebo dokumenty bez interakcie s mobilom. Obrazovku smartfónu je možné zdieľať na obrazovke počítača a vzdialene ovládať smartfón napríklad cez TeamViever. Vďaka spoločnosti ifunplay a aplikácii DocScan možno teraz ScanTent používať aj s operačným systémom iOS v iPhonoch. Držiak na vrchnej časti zariadenia ScanTent umožňuje umiestnenie smartfónu, optimálny pozorovací uhol a konštantnú vzdialenosť. Ak denné svetlo nestačí, biele LED pásiky poskytujú rovnomerné osvetlenie, ktoré maximalizuje kvalitu obrazu.

Dokument (*Document*). V štruktúre systému Transkribus expert klient je dokument zvyčajne zaradený do zbierky. Dokument môže byť presunutý do inej existujúcej zbierky. Základné metadáta k dokumentu sú: jedinečný číselný identifikátor, názov dokumentu, meno osoby, ktorá nahrala dokument do zbierky v Transkribe, dátum a čas nahratia do zbierky, meno zbierky, do

ktorej dokument patrí. Dokument je možné zobraziť vo forme Prehľad (*Overview*) s jednotlivými stranami a grafickým rozlíšením stavu stránky (napr. *Ground Truth*, *In progress*, *Done*, *Final*). Vo forme Rozloženie (*Layout*) sú viditeľné texty transkripcie strán, riadky textu, poradie čítania riadkov strojom, identifikátor riadka a koordináty umiestnenia elementov v riadku.

Export. Ak chceme pracovať s obrázkami a prepismi mimo Transkribu, môžeme svoje dokumenty exportovať do bežnejších formátov, ako sú docx, PDF, xls, PageXML, TEI-XML alebo txt. Možnosti zahŕňajú export celých strán, obrázkov, textu alebo štrukturálnych prvkov. Exportovať je možné do adresára na lokálnom počítači alebo exportovať na server Transkribus, z ktorého príde oznámenie po skončení exportu.

Formát JPG, JPEG. Najrozšírenejší je formát, ktorý sa vyskytuje s príponou .jpg, .jpeg. V tomto formáte ukladajú súbory všetky fotoaparáty aj mobilné zariadenia, ak používame napríklad DocScan. V niektorých aparátoch je možné voliť jeden formát alebo snímanie v dvoch formátoch JPG a RAW (ARW). Výhodou formátu JPG je, že sa obrázok dá zobraziť prakticky v každom zariadení – v mobilnom telefóne, televízore alebo vo webovom prehliadači. Zaberá málo miesta na disku, je úsporný, pretože ide o kompresiu so stratou. Nevýhodou tohto formátu je, že každou úpravou obrázok stráca kvalitu pri každom uložení. V projektoch transkripcie používame na snímanie mobilnými zariadeniami formát JPG na archivovanie a v transkripcii spravidla pracujeme s derivovaným formátom PDF.

Formát PNG. Skratka v preklade znamená prenosná sieťová grafika (*Portable Network Graphics*), čiže ide o bezstratový kompresný formát pre obrázky a fotografie využívaný najmä na internete.

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

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

Formáty obrázkov. Snímky je možné tvoriť, ukladať a upravovať v rôznych formátoch. Najčastejšie ide o súbory vo formátoch RAW a JPG. Z hľadiska úprav fotografií je dôležitý formát TIFF.

Gotické písmo malo niekoľko druhov. Napríklad francúzska textúra s veľmi ostrým lomom a štíhlou stavbou, talianska širšia a okrúhlejšia rotunda s miernejším lomením oblúkov, zmiešané písmo – bastarda, v Nemecku švabach – písmo širších, oválnejších tvarov a fraktúra – písmo užších a špicatejších tvarov s ozdobnými úponkami. Vynálezom kníhtlače (v roku 1450 Johannom Gutenbergom) sa tento druh písma veľmi rozšíril najmä v krajinách hovoriacich po nemecky.

Ground Truth (základná pravda) je vzorka manuálne prepísaných a dôsledne skontrolovaných a korigovaných strán dokumentu používaná pri trénovaní modelu automatickej transkripcie.

HTR+ a **PyLaia**. Softvér HTR+ spoločnosti Transkribus zatiaľ nemôže okamžite spustiť spoľahlivý automatický prepis, ale najprv musí byť vyškolený na konkrétny typ písma a rukopisu. HTR+ vyvinutý tímom CITlab na Univerzite v Rostocku bol do konca roka 2022 aj vo výskume Skriptor používaný ako hlavný stroj na rozpoznávanie rukopisného textu. Transkripčný me-

chanizmus je založený na TensorFlow. Namiesto HTR+ je v súčasnosti v Transkribe dostupný nástroj PyLaia.

Import dokumentov (*Upload*). Po vytvorení zbierky v Transkribe je potrebné nahrať dokumenty. Potom je možné spustiť nástroje, ako sú analýza rozloženia (segmentácia) alebo rozpoznávanie textu (transkripcia). Údaje v Transkribe sú vždy súkromné a prístupné iba jednotlivým používateľom. Vlastník zbierky (*Owner*) môže umožniť prácu aj iným používateľom (*Users*) s oprávneniami, ktoré im pridelí (*Owner*, *Editor*, *Transcriber*, *Reader*).

ISAD(G) (*General International Standard Archival Description*). Medzinárodný štandard, ktorý definuje zoznam prvkov a pravidiel na popis archívov a popisuje druhy informácií, ktoré musia a mali by byť zahrnuté v takýchto opisoch. Vytvára hierarchiu popisu, ktorá určuje, aké informácie by mali byť zahrnuté na akej úrovni. V súvislosti s výskumom a experimentmi s transkripciou archívnych dokumentov považujeme za vhodné, aby boli transkribované fondy, zbierky a dokumenty popísané na štandardnej úrovni. Tento štandard poskytuje rámec pre spoločný prístup a nie rigidný formát.

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

OCR (*Optical Character Recognition*). Optické rozpoznávanie znakov alebo optická čítačka znakov je elektronická alebo mechanická konverzia obrázkov ručne písaného alebo vytlačeného textu na strojovo kódovaný text či už z naskenovaného dokumentu alebo fotografie.

Overovací súbor (*Validation Set*) pozostáva zo strán dokumentu, na ktorých sa presnosť vytrénovaného modelu automaticky overí (odskúša). V porovnaní s cvičným súborom je preto menší, spravidla 10 % z celkovej vzorky *Ground Truth*. Na druhej strane overovací súbor by mal byť reprezentatívny, t. j. mal by obsiahnuť príklady všetkých písmen, jazykov a iných atribútov zahrnutých v cvičnom súbore. V opačnom prípade, čiže ak je overovací súbor príliš homogénny, výkon modelu môže byť nízky, prípadne skreslený.

Polygóny (*Polygons*). Historické dokumenty majú niekedy zložité usporiadanie a pozostávajú z rôznych rozložení, čo môže viesť k problémom s poradím čítania prvkov textu. Pri komplikovaných rozloženiach si rýchlo všimneme, že ručne nakreslené textové oblasti sa môžu prekrývať. Tento problém sa dá ľahko vyriešiť úpravou pravouhlých oblastí textu, pridaním bodov a tým vytvorením polygónov.

Poradie čítania. V systéme Transkribus expert klient poradie čítania zobrazuje na segmentovanej stránke to poradie, v ktorom bude stroj transkripcie čítať riadky textu na obrázku stránky. Toto poradie čítania sa vytvára automaticky počas segmentácie, ale možno ho neskôr zmeniť aj manuálne. Pri automatickej analýze rozloženia je poradie čítania určené súradnicami riadkov na obrázku: horný riadok, ktorý je najviac vľavo, je číslo jedna atď. Dôležité je vedieť, že poradie čítania nie je relevantné pre samotné trénovanie, ale môže sťažovať čítanie transkribovanej strany. Ak sa má prepis exportovať a ďalej použiť na vydanie, tak poradie čítania je potrebné zadať správnym spôsobom, aby bol text v správnom poradí. Dá sa to jednoducho urobiť zapnutím poradia čítania ikonkou Viditeľnosť tvaru (*Shape visibility*). Vo všetkých riadkoch sa tak zobrazí krúžok s číslom, ktoré označuje ich polohu na stránke dokumentu. Kliknutím na tieto

krúžky sa otvorí okno s textovým editorom, kde je možné priradiť nové, správne čísla. Táto funkcia je užitočná najmä v dokumentoch s náročným rozložením, kde sa poradie riadkov neriadi bežnými pravidlami.

Presnosť modelu. Presnosť modelu je možné merať na konkrétnych stránkach z cvičných a overovacích súborov pomocou funkcie Porovnať (*Compare...*) na záložke Nástroje (*Tools*). Na tento účel je najprv potrebné generovať automatický prepis. Na porovnanie textových verzií sú potrebné dva transkribované súbory: referencia (*Reference*) – správny text a hypotéza (automaticky transkribovaný text). Ako referencia sa vyberie verzia stránky, ktorá bola správne prepísaná, teda „základná pravda“ (*Ground Truth*), čo je manuálny prepis čo najbližšie k pôvodnému textu. Na získanie najvýznamnejšej hodnoty by bolo najlepšie použiť stránky zo vzorového súboru, ktoré neboli použité v tréningu, a preto sú pre model nové. Použitie stránok z overovacieho súboru je tiež možnosťou, aj keď nie ideálnou. Použitie stránok z cvičného súboru nie je vhodné, pretože to prinesie nižšie hodnoty CER, ako v skutočnosti sú. Ako hypotézu vyberieme verziu, ktorá bola automaticky vygenerovaná pomocou vytrénovaného modelu, a na ktorej chceme vidieť, aký dobrý je výsledok.

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

PyLaia. Nástroj na rozpoznávanie rukopisného textu, ktorý umožňuje používateľovi nastaviť si jednotlivé parametre transkripcie. Zmeniť sa dá aj sieťová štruktúra PyLaia, čo je príležitosť pre ľudí, ktorí poznajú strojové učenie. Úpravy neurónovej siete je možné vykonať prostredníctvom úložiska GitHub. Dokumenty, ktoré boli transkribované pomocou modelu PyLaia, je možné prehľadávať pomocou plnotextového vyhľadávania (*Solr*) v Transkribe.

READ (*Recognition and Enrichment of Archival Documents*). Projekt, ktorého riešenie prebiehalo v rokoch 2016 – 2019 v rámci programu Horizon2020. Výskum bol predtým financovaný ako súčasť projektu *tranScriptorium*. Tento projekt získal finančné prostriedky zo 7. rámcového programu Európskej únie pre výskum, technologický rozvoj podľa dohody o grante č. 600707. Viac o projekte https://cordis.europa.eu/project/id/674943

Read&search. Platforma Transkribu, ktorá sprístupňuje dokumenty zo zbierky vytvorenej v platforme Transkribus expert klient online formou. Webové rozhranie bohaté na funkcie je ideálne na sprístupnenie historických dokumentov a vyhľadávanie na webe.

READ-COOP. Združenie na udržateľnosť a vývoj platformy Transkribus. V októbri 2022 malo združenie 113 členov z 27 krajín. Jedinou členskou krajinou zo strednej a východnej Európy bolo v tom čase Slovensko. V READ-COOP sa kupujú kredity. Nejde o zisk združenia, ale o príjem, ktorý sa používa na výskum, vývoj a infraštruktúru platformy.

Riadkové rámce (*Line Regions, LR*). Oblasti, ktoré sa nachádzajú v textových rámcoch a možno ich opísať ako mnohouholníky, v ktorých je všetok ručne písaný/tlačený text v riadku. Keďže nemajú pre proces transkripcie bezprostredný význam, riadkové rámce by sa nemali opravovať. Ak sa niečo má zmeniť v rozložení riadkov dokumentu, vždy to treba urobiť na úrovni základ-

nej čiary (*Baseline*). Základná čiara by mala prebiehať pozdĺž spodnej časti textového riadku, písmená by na nej mali sedieť a zostupne smerovať nižšie. Riadkové rámce sa prispôsobia automaticky, keď niečo zmeníte na základnej úrovni. Zobrazí sa vyskakovacie okno s otázkou, či chcete zmeniť aj nadradený riadok, čo treba potvrdiť.

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

SKRIPTOR. Projekt APVV-19-NEWPROJECT-17816 (2020 – 2024). Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov (*Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts*). Riešiteľské organizácie: Univerzita Mateja Bela v Banskej Bystrici (zodpovedný riešiteľ doc. Imrich Nagy, PhD.), Štátna vedecká knižnica v Banskej Bystrici – partner (garant prof. PhDr. Dušan Katuščák, PhD.).

Snímanie je jeden z procesov digitalizácie. Vykonáva sa pomocou vhodného technického zariadenia na digitalizáciu, akými sú zariadenia na zachytenie digitálneho obrazu (digitálne fotoaparáty a kamery, skenery na knihy, dokumenty alebo mikrofilmy, audio- a videohardvér) pripojené na vhodnú počítačovú platformu. Je možné rozlíšiť dve rôzne metódy snímania: skenovanie a fotografovanie, používanie digitálnych kamier/fotoaparátov, mobilných telefónov. Na účely automatickej transkripcie, pokiaľ je to možné, použijeme dokumenty nasnímané profesionálnymi skenermi a obrazmi v najvyššej dosiahnuteľnej kvalite. Minimálna kvalita skenovania by mala byť 300 DPI. Nakoľko pri historických rukopisoch ide de facto o grafiku, je vhodné skenovať vo vyššej kvalite. Pre platformu Transkribus je možné snímať dokumenty do formátu veľkosti A3 zariadením ScanTent so softvérom DocScan.

Stav dokumentu. Rôzne stavy spracovania strany: *New* (nový – stav pre novonahraté dokumenty), *In Progress* (prebiehajúci – automatická zmena stavu po úprave strany), *Done* (hotový – stránka je prepísaná, ale vyžaduje ešte ďalšiu kontrolu), *Final* (finálna verzia – stránka prepísaná a skontrolovaná), *Ground Truth* (základná pravda – 100 % správne prepísaná strana). Znamená to, že sa zaznamenáva práca s každou jednotlivou stranou a verzii strany sa môžu priradiť rôzne stavy v závislosti od toho, aký pokrok sa na nich dosiahol.

Štrukturálne metadáta (tagy) (*Structural metadata – tags*). V štruktúre systému Transkribus expert klient je možné pomocou funkcie štrukturálneho značkovania vo funkcionalite metadáta označiť, „značkovať“ (*Mark-up*) prvky štruktúry dokumentov. Okrem toho je možné trénovať modely tak, aby automaticky rozpoznali štruktúru dokumentov. Pridaním tagov, teda štrukturálnych značiek sa vytvoria cvičné dáta pre tento proces. Nie je potrebné označovať každý prvok dokumentu, stačí sa zamerať na označenie sekcií, ktoré nás zaujímajú. Rozhranie štrukturálneho označovania v Transkribe umožňuje rozdeliť dokumenty do štruktúrnych sekcií, ako sú odseky, nadpisy alebo čísla strán, pridať prispôsobené kategórie značiek pre vaše individuálne potreby a v budúcnosti použiť tieto štrukturálne informácie na trénovanie modelu.

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

Textový rámec (*Text region, TR*). Ak chceme vygenerovať automatický prepis pomocou platformy Transkribus, musíme dokumenty rozdeliť na textové rámce, v nich vymedziť riadkové rámce a základné čiary. V predvolenom nastavení je oblasť textu obdĺžnik, ktorý obklopuje všetok ručne písaný text obsiahnutý v obrázku. Je však možné upraviť textový rámec podľa všeobecného rozloženia pridaním kontrolných bodov, čím sa vytvorí polygón.

Transkribus expert klient (*Transkribus Expert Client*). Samostatná profesionálna verzia Transkribu s plným výkonom platformy Transkribus. Posledná verzia je verzia 1.26.0 z 5. júna 2023. V súčasnosti sa vývojový tím aplikácie sústreďuje výlučne na zdokonaľovanie a ďalší vývoj webového rozhrania Transkribus web app (*Transkribus Lite*).

Transkribus web app (*Transkribus Lite*). Online verzia aplikácie pre platformu Transkribus k 18. októbru 2023 vo verzii 3.0.1.26. Automaticky transkribuje a umožňuje pohodlnú úpravu historických dokumentov. V súčasnosti už má aplikovanú väčšinu funkcionalít Transkribus expert klienta. V Transkribus web app je teda možné realizovať všetky fázy potrebné na automatickú transkripciu: import dokumentu, segmentáciu, trénovanie modelu, automatickú transkripciu a export transkripcie vo zvolenom formáte.

Transkribus. Komplexná platforma na digitalizáciu, na rozpoznávanie textu podporované umelou inteligenciou, ako aj na prepis a vyhľadávanie historických dokumentov – z akéhokoľvek miesta, kedykoľvek a v akomkoľvek jazyku. Platforma integruje nástroje vyvinuté výskumnými skupinami v celej Európe vrátane skupiny na rozpoznávanie vzorov a technológie ľudského jazyka Technickej univerzity vo Valencii a skupiny CITlab University Rostock. V októbri 2023 mal Transkribus viac ako 100 000 registrovaných používateľov a viac ako 40 miliónov rozpo-

znaných strán. Platforma bola vytvorená v kontexte dvoch projektov EÚ *tranScriptorium* (2013 – 2015) a *READ* (2016 – 2019).

Transkripcia (prepis). Na platforme Transkribus sa používa termín transkripcia vo význame prepisu rukopisného alebo tlačeného historického textu v určitom jazyku a automatický prepis textu v tom istom jazyku. Napríklad rukopis v maďarčine sa prepisuje pomocou znakovej sady tlačenej latinky. Nejde teda o prepis medzi jazykmi, ale o prepis v rámci jedného jazyka.

Transliterácia. Ortograficky vernému prepisu zodpovedá označenie transliterácia. Na platforme Transkribus sa pre všetky druhy prepisu konvenčne používa pojem transkripcia.

Trénovanie modelu. Pomocou nástroja Transkribus expert klient je možné trénovať model rozpoznávania rukopisného textu, aby bolo možné automaticky transkribovať zbierky dokumentov. Model je výsledkom trénovania, preto je pri jeho tvorbe potrebné trénovať tak, aby stroj rozpoznal určitý štýl písania v zobrazovaných obrázkoch dokumentov a poskytol ich viac-menej presný prepis. Na trénovanie modelu je potrebných 5 000 až 15 000 slov (približne 25 – 75 strán) prepísaného materiálu. Prepis sa získa manuálnym prepisom riadok po riadku presne podľa predlohy. Prepis si možno uľahčiť použitím už prepísaných a dostupných dokumentov alebo postupovať pri príprave cvičného súboru s použitím základného súboru. Pri práci s tlačeným textom sa zvyčajne vyžaduje menšie množstvo cvičných údajov ako pri rukopisoch. Použitím základného modelu je možné znížiť množstvo požadovaných cvičných dát. Ako základný model sa môže použiť buď jeden z verejne dostupných modelov PyLaia v Transkribe, ktorý by mohol byť vhodný pre naše dokumenty, alebo jeden z našich vlastných modelov, ktoré sme už predtým cvičili.

Verejné modely transkripcie (*Public Models*) sú modely Transkribu, ktoré je možné použiť na podobné dokumenty. Pre každý model je uvedený krátky opis cvičného materiálu, pre ktoré jazyky môže byť model užitočný a kto ho vytvoril a cvičil. Cieľom je sprístupniť používateľom Transkribu čoraz viac modelov, aby mohli ťažiť z kooperácie a sieťového efektu, a šetriť prácu a čas. V súčasnosti je dostupných viac ako 100 verejných modelov napríklad: nemecký kurent, noviny, časopisy, rôzne tlače a rukopisy; viacjazyčný model pre tlače v rôznych jazykoch (holandčina, angličtina, fínčina, francúzština, nemčina, švédčina); všeobecný model pre francúzske rukopisy, nemecká bastarda 15. st.; dánska fraktúra a historické rukopisy a strojopisy; holandské rukopisy a tlače; estónske rukopisy; fínske noviny a rukopisy; francúzske rukopisy a tlače; hlaholika; latinčina; neolatinčina; ruština; španielske rukopisy a tlače a i.

Verzie. Pri práci so systémom Transkribus expert klient sa pri každom spustení úlohy alebo uložení dokumentu vytvorí nová verzia dokumentu. Výhodou je, že sa vždy môžete vrátiť k starším verziám a pokračovať v práci na nich, čo zabraňuje strate údajov v Transkribe. Verzie je možné porovnávať pomocou funkcie Porovnať (*Compare*). Pri verziách jednotlivých stránok je vždy informácia o stave strany (*Page status*), používateľovi, dátume zmeny, nástroji zmeny a identifikátoroch.

Virtuálna klávesnica. Editačný nástroj Transkribus expert klient, ktorý umožňuje pridávať znaky sady Unicode (ISO 10646) a špeciálne znaky, ktoré nie sú dostupné na bežnej klávesnici. Nachádza sa v poli textového editora v spodnej časti okna expertného klienta. Pomocou tlačidla Upraviť (*Edit*) je možné pridávať skratky pre často používané znaky a pridávať nové znaky Unicode. Ak je potrebné vytvoriť skratku, stačí ju zadať do stĺpca skratka a na pridanie nových znakov Unicode použiť zelené tlačidlo plus.

WER (*Word Error Rate*). Hodnota chybovosti slov v transkripcii.

Základná čiara (*Baseline, BL*). Najdôležitejší referenčný bod na rozpoznávanie textu. Popisuje polyčiaru, ktorá sa tiahne pozdĺž spodnej časti rukou písaného/tlačeného textového riadku. Segmentáciu textu na riadkové rámce a základné čiary je možné vykonať automaticky pomocou Transkribus LA. Pri zložitých rozloženiach a v závislosti od konkrétneho písma v rukopisoch/tlačiach sa však môžu vyskytnúť prípady, keď je potrebné vykonať niektoré manuálne opravy. Základná čiara by mala prebiehať pozdĺž spodnej časti textového riadku, písmená by na nej mali sedieť a zostupne smerovať nižšie. Základná čiara pozostáva z jednotlivých bodov, ktoré je možné nastaviť pri manuálnej úprave segmentácie.

Základný model (*Base model*). Ak tvoríme vlastné, generické modely HTR, tak nepracujeme so základnými modelmi. Pri trénovaní so základnými modelmi je však každé trénovanie pre model založené na existujúcom modeli, t. j. na základnom modeli. Toto je spravidla posledný model HTR, ktorý bol vytrénovaný v nejakom projekte. Základné modely si „pamätajú" to, čo sa už „naučili". Preto každé nové trénovanie teoreticky zlepšuje kvalitu novotvoreného modelu. Nový model sa učí od svojho predchodcu a stáva sa tak lepším a lepším. Preto je trénovanie so základnými modelmi obzvlášť vhodné aj pre veľké generické modely, ktoré sa neustále vyvíjajú počas dlhého časového obdobia. Ak chceme vykonať trénovanie so základným modelom, jednoducho si v cvičnom nástroji okrem obvyklých nastavení vyberieme konkrétny základný model. Potom na karte údaje modelu HTR (*Model data*) vložíme cvičný súbor a overovací súbor základného modelu, ako aj nový cvičný a overovací súbor. Okrem toho môžeme pridať ďalšie nové strany *Ground Truth* a začať s cvičením.

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

Zbierka (*Collection*). V štruktúre systému Transkribus expert klient sú dva kľúčové prvky: zbierky a dokumenty. Zbierka je nadradená dokumentu. Dokumenty sú usporiadané do zbierok. Zbierky možno chápať ako priečinky obsahujúce dokumenty. Zbierky sa zvyčajne tvoria podľa konkrétneho projektu. Napríklad všetky dokumenty patriace k jednému projektu sú usporiadané do jednej zbierky. Dokumenty pozostávajú z jednej alebo viacerých strán dokumentu. Každá zbierka v Transkribe má jedinečný identifikátor (ID). Každý dokument v zbierke má jedinečný číselný identifikátor, názov dokumentu, počet strán dokumentu, meno osoby, ktorá nahrala dokument do Transkribu, dátum a čas nahratia, meno vlastníka zbierky. V zbierke je možné manažovať – tvoriť, vymazať, upravovať, pridávať a upravovať oprávnenia používateľom zbierky so súhlasom a rozhodnutím vlastníka zbierky, pracovať s kreditmi k zbierke. Ku každému dokumentu je možné popísať všeobecné metadáta a metadáta k jednotlivým stranám, ako aj štrukturálne a textové metadáta a komentáre. Používateľ môže mať niekoľko zbierok s rôznymi dokumentmi. Na účely prezentačnej vrstvy *Read&search* je potrebné vytvoriť jednu spoločnú zbierku. Všetky zbierky a dokumenty v Transkribe sú súkromné.

Použité zdroje

DRAŠKABA, Peter a Jozef HANUS, prekl. Všeobecná medzinárodná norma pre opis archívnej jednotky. *Slovenská archivistika* [online]. 2000, roč. 35, č. 1, s. 197 – 215 [cit. 2023-08-17]. ISSN 2730-0323. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/Slovenska%20archivistika_1-2020.pdf

KATUŠČÁK, Dušan: Metodológia a metodika transkripcie historických textov. In: KATUŠČÁK, Dušan a Imrich NAGY, eds. *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela, 2022, s. 18 – 47 [cit. 2023-08-29]. ISBN 978-80-557-2020-3. Dostupné na: https://doi.org/10.24040/2022.9788055720203

KERESTEŠ, Peter. Archívny dokument a jeho definícia. *Slovenská archivistika* [online]. 2022, roč. 52, č. 1, s. 137 – 147 [cit. 2023-18-17]. ISSN 2730-0323. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/SA%201-2022,%20roc.%2052.pdf

KÖRMENDY, Lajos. Štandardizovanie opisu archívnej jednotky: odborný nástroj v kontexte národnej a regionálnej tradície. *Slovenská archivistika* [online]. 2000, roč. 35, č. 2, s. 222 – 235 [cit. 2023-08-17]. ISSN 2730-0323. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/Slovenska%20archivistika_2-2020.pdf

KURHAJCOVÁ, Alica. Keď sa stroj učí čítať Hurbanove listy. In: *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela, 2022, s. 124 – 145 [cit. 2023-10-09]. ISBN 978-80-557-2020-3. Dostupné na: https://doi.org/10.24040/2022.9788055720203

Metodický pokyn odboru archívov sekcie verejnej správy Ministerstva vnútra SR o postupe štátnych archívov pri digitalizácii archívnych dokumentov a tvorby povinných metadát č. SVS-OA-2011/23406-001 [online]. Bratislava, 2011 [cit. 2023-08-18]. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/odbor_archivov_a_registratur/MP_digitalizaciaAD_metadata.pdf

NAGY, Imrich. Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov. *Slovenská archivistika* [online]. 2021, roč. 51, č. 2, s. 53 – 67. Dostupné na: https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/SA%202-2021,%20roc.%2051.pdf

NAGY, Imrich. Sprístupnenie Csákósovho katalógu korešpondencie Koháryovcov pomocou automatickej transkripcie. In: KATUŠČÁK, Dušan a Imrich NAGY, eds. *Automatická transkripcia slovacikálnych historických dokumentov* [online]. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela, 2022, s. 66 – 83 [cit. 2023-08-15]. ISBN 978-80-557-2020-3. Dostupné na: https://doi.org/10.24040/2022.9788055720203

PÉKOVÁ, Monika. Od analógového archívneho dokumentu k jeho digitálnej kópii. In: GRESCHOVÁ, Eva a František CHUDJÁK, eds. *Zborník Spoločnosti slovenských archivárov 2015.* Bratislava: Spoločnosť slovenských archivárov, Slovenské múzeum ochrany prírody a jaskyniarstva, 2016, s. 78 – 81. ISBN 978-80- 971356-2-1.

Resource Center. In: *READ-COOP* [online]. Innsbruck: READ-COOP SCE, last update 2023 [cit. 2023-08-28]. Dostupné na: https://readcoop.eu/transkribus/resources/

ŠEDIVÝ, Juraj a Hana PÁTKOVÁ, eds. *Vocabularium parvum scripturae latinae* [online]. Bratislava – Praha, 2008 [cit. 2023-08-25]. Dostupné na: https://manuscripta.at/Ma-zu-Bu/dateien/Vocabularium_parvum_scripturae_Latinae_2008.pdf

Transkribus: help center [online]. [cit. 2023-08-28]. Dostupné na: https://help.transkribus.org/

Všeobecný medzinárodný štandard pre archívny opis ISAD(G) [online]. 2. vyd. Bratislava, 2015 [cit. 2023-08-18]. Dostupné na: https://www.minv.sk/?archivne-standardy-1

DOI: https://doi.org/10.24040/2023.9788055720708

ISBN 978-80-557-2070-8