RECENZOVANÉ PŘÍSPĚVKY

Pokrok v transkripci historických rukopisných dokumentů[1]

Progress in the Transcription of Historical Manuscript Documents

prof. PhDr. Dušan Katuščák, PhD. (ORCID 0000-0001-7444-1077), Mgr. Klára Pohlová, Bc. Lukáš Němec, BcA. et Bc. Vojtěch Říha / *Slezská univerzita v Opavě, Filozoficko-přírodovědecká fakulta, Ústav bohemistiky a knihovnictví (Silesian University in Opava, Faculty of Philosophy and Science, Institute of the Czech Language and Library Science), Masarykova třída 343/37, 746 01 Opava*

RESUMÉ: Studie je zaměřena na pokrok v transkripci historického písemného dědictví v Česku a na Slovensku od roku 2020. Odkazuje na výzkumné aktivity, experimenty a výsledky dosažené v letech 2020–2024 v kontextu platformy Transkribus v projektu SKRIPTOR[2]. Zmiňuje se o českých výzkumných projektech Vysokého učení technického (VUT) v Brně, jejichž výsledkem je nástroj pro transkripci PERO. Informuje o nejnovějších modelech transkripce v platformě Transkribus. Těžiště studie spočívá v popisu postupu a experimentů při tvorbě modelů deseti českých rozličných historických rukopisných dokumentů, které v rámci projektu Studentské grantové soutěže SGS 2024 na Slezské univerzitě v Opavě provedli studenti Lukáš Němec a Vojtěch Říha. Do studie je zařazen i stručný popis tvorby modelu transkripce strojopisných dokumentů, který vytvořila jako součást projektu SGS 2023 Klára Pohlová.

KLÍČOVÁ SLOVA: modely transkripce, historické rukopisy, transkripce českých dokumentů, transkripce slovenských dokumentů, platforma Transkribus

SUMMARY: The study focuses on the progress in the transcription of historical written heritage in the Czech Republic and Slovakia since 2020. It highlights research activities, experiments and results achieved between 2020 and 2024 in the context of the Transkribus platform within the SKRIPTOR project. It also mentions Czech research projects from the Brno University of Technology, which resulted in the PERO transcription tool. In addition, it provides information about the latest transcription models available on the Transkribus platform. The study also describes the procedures and experiments in creating models of ten different Czech historical manuscript documents, which were carried out by students Lukáš Němec and Vojtěch Říha as part of the 2024 Student Grant Competition project at the Silesian University in Opava. The study also includes a brief description of the development of a transcription model for typewritten documents, created by Klára Pohlová as part of the SGS 2023 project.

KEYWORDS: transcription models, historical manuscripts, transcription of Czech documents, transcription of Slovak documents, Transkribus platform

1 Studie vznikla díky projektu Studentské grantové soutěže (SGS) 2024 na Slezské univerzitě v Opavě, Filozoficko-přírodovědecké fakultě, Ústavu bohemistiky a knihovnictví, Oddělení knihovnictví.

2 Odkaz na článek o projektu SKRIPTOR: *https://knihovnarevue.nkp.cz/archiv/2022-2/recenzovane-prispevky/umela-inteligencia-pomaha-spristupnovat-pisomne-dedicstvo*.

1 Úvod (Dušan Katuščák)

Staré a vzácné tisky, strojopisy, a hlavně rukopisy zpravidla nelze uspokojivě transkribovat pomocí nástrojů *optického rozpoznávání písma* (*OCR*). Přichází na pomoc *umělá inteligence*. Ve snahách zpřístupnit historické písemné dědictví z digitálních repozitářů se pozornost výzkumníků koncentruje na transkripci a strojové učení s použitím konvolučních *neuronových sítí*. Jedná se o proces, ve kterém se pořízený obrázek „mění" na text. Tedy *pixely* se „mění" na *byty* (*bajty*). V posledních pěti letech se k transkripci používají různé *platformy* a nástroje open source i komerčně zaměřené *nástroje* a *služby*. Náš zájem o problematiku transkripce byl podnícen vědeckým evropským projektem základního výzkumu READ, který se realizoval díky programu Horizon 2020. Autorem a koordinátorem projektu byl prof. G. Mühlberger z Univerzity v Innsbrucku. Projekt READ byl financován Evropskou unií částkou 8,2 milionu EUR. Financování skončilo 30. 6. 2019. V současnosti projekt pokračuje na bázi sdružení READ-COOP (READ, 2024). Začátkem roku 2024 měly aplikace Transkribus přes 400 interaktivních uživatelů denně. Uživatelé do systému nahráli denně v průměru 25 tisíc digitalizátů a vytvořili 15 modelů pro rozpoznávání textu. Statistiky uvádějí, že od roku 2015, kdy platforma začala působit, bylo zpracováno přes 51,5 milionu digitálních faksimilií a vytvořeno cca 25 560 modelů, na kterých pracovalo přes 171 307 lidí na celém světě (Nockels et al., 2024).

Jelikož jsem byl jedním ze tří hodnotitelů projektu READ pro Evropskou komisi, chtěl jsem vědět, co posuzuji. Začal jsem se proto o problematiku transkripce podrobněji zajímat. Z praxe digitalizace jsem věděl, že zatímco optické rozlišení tištěného písma (OCR) v procesu digitalizace dostatečně zvládá například vynikající nástroj OCR *ABBY FineReader*[3], pak rozpoznávání textů v historických tištěných dokumentech, rukopisech a strojopisech je nedostatečné a výsledky transkripce jsou neuspokojivé. Sám jsem od roku 2018 věnoval tisíce hodin experimentům a tvorbě modelů v platformě Transkribus. Zpočátku to byl entuziasmus a osobní iniciativa. O výsledcích jsem informoval odbornou veřejnost v různých prezentacích, zvaných přednáškách a publikacích (Katuščák, 2020a; Katuščák, 2020b; Katuščák, 2022a).

V roce 2020 jsem inicioval projekt SKRIPTOR (Katuščák, 2022b). Díky porozumění historiků a archivářů z *Katedry historie na Univerzitě Mateje Bela v Banské Bystrici* a zvláště *doc. Imricha Nagye* jsme podali projekt a získali jsme podporu 170 000 eur z Agentury na podporu vědy a výzkumu pro projekt, který se realizoval v letech 2020–2024. Naše úsilí jsme koncentrovali na zvládnutí platformy Transkribus a tvorbu modelů transkripce. Autorské privátní modely výzkumníků jsme nakonec zpracovali v agregovaných supermodelech pro transkripci historických *rukopisů* s chybovostí CER[4] 5,30 % (SUPERMODEL_M1, 2024). Tento model úspěšně ověřil Imrich Nagy na transkripci latinského historického rukopisu *Acta comitatus Nitriensis sedis iudiciariae* s mírou chybovosti jen 2,20 % (Nagy, 2024). Pro transkripci *historických tisků* a *strojopisů* byl vyvinut původní supermodel (SUPERMODELP&T1, 2024) s chybovostí 1 %.

---

3 Odkaz na webové stránky OCR ABBY FineReader: *https://pdf.abbyy.com/*

4 CER (Character Error Rates). Míra chybovosti znaků (srovnává pro danou stranu celkový počet znaků (n) včetně mezer s minimálním počtem vložení (i), nahrazení (s) a vymazání (d) znaků, které jsou potřebné k získání výsledku Ground Truth. Jedná se tedy o chyby v porovnání s přesným, referenčním textem. Vzorec pro výpočet CER je následující: CER = [(i + s + d) / n]*100. Každá malá chyba v přepisu je statisticky plnohodnotná chyba. Obecně lze konstatovat, že: a) je-li hodnota chybovosti znaků CER nižší než 10 %, což je 10 a méně chyb na sto znaků, tak výsledek transkripce je dobrý, čitelný a, je-li to účelné, je možné další editování výstupu; b) je-li míra chyb znaků CER ≤ 5 %, je výsledek transkripce velmi dobrý; c) je-li míra chyb znaků CER nižší než 3 %, lze výsledky transkripce považovat za vynikající a míra chyb znaků CER nižší než 2,5 % za excelentní.

Ve výzkumu SKRIPTOR, v projektech SGS (Katuščák, 2024) a diplomových pracích (Smida, 2023; Pohlová, 2024) preferujeme platformu Transkribus, kterou osobně považuji za bezkonkurenčně nejlepší na světě.

V Česku dominuje ambiciózní nástroj PERO (Žabička, 2023; Zavřelová, 2020), který se vyvíjel v rámci výzkumu na VUT v Brně pod vedením Michala Hradiše (Hradiš et al., 2024) v letech 2018–2022. Tým poskytuje volně dostupný nástroj transkripce i komerční služby transkripce. Nejnovější *OCR motory* jsou dostupné na *pero-ocr.fit.vutbr.cz*. *OCR motory* jsou dostupné také přes API *spuštěné* na *pero-ocr.fit.vutbr.cz/api*, *github repository*.

Existují i jiné nástroje transkripce, nicméně pro jejich důkladné srovnání a ohodnocení je nutná *metaanalýza* s jasně stanovenými kritérii hodnocení a následnou identifikací *nejlepší dostupné technologie*. Taková metaanalýza však není předmětem této studie.

Základem automatické transkripce jsou kvalitní *modely*. Jen platformy a nástroje, které mají zabudované dobré modely, jsou schopny produkovat přijatelné až excelentní výsledky transkripce s chybovostí pod 8–5 % CER, čehož lze dosáhnout pouze provedením množství experimentů, zkoušení, nastavování parametrů segmentace textu apod. (Katuščák et al., 2023). Modely tedy slouží k transkripci historických textů, přičemž se pro tvorbu modelů využívá umělá inteligence. Pro vytvoření modelu je třeba stroj *naučit*, co má dělat. *Strojové učení* probíhá tak, že se manuálně připraví *tréninkový set* (Train set)[5] a *validační set* (Validation set). Strany textu je třeba ručně přepsat co nejpřesněji do kvality *GT* (Ground Truth)[6]. Následně se spustí proces trénování, cvičení stroje. Výsledkem trénování je *MODEL*. Na základě dílčích modelů lze pak připravit univerzální supermodely.

V platformě Transkribus bylo v roce 2024 k dispozici mnoho *veřejně dostupných* (237) a *privátních* (275) modelů, které však zatím nejsou vhodné pro západoslovanské jazyky, resp. texty naší provenience. K vytvoření těchto supermodelů byly zapotřebí pro trénování miliony slov. K dispozici je například model *Titan* (TITAN, 2023) pro německé, anglické, holandské, francouzské, finské a švédské rukopisné texty 16.–20. století. Existuje také velmi kvalitní model pro transkripci němčiny *The German Giant* (GIANT, 2023) vytvořený na základě 86 345 stran a 15 420 976 slov. Efektivnost transkripce německého rukopisu jsme ověřili v projektu SGS v roce 2022. Předmětem transkripce byla německá rukopisná kuchařská kniha z roku 1667 (KACH, 1667) o 876 stranách. Je zřejmé, že pokud chceme mít pro transkripci historických rukopisů západoslovanské provenience (bohemika, slovacika, polonika...) použitelný nástroj, čeká nás množství trpělivé práce na tvorbě vlastních modelů, které se stanou součástí větších (GIANT, 2023) supermodelů.

Usilovali jsme o přenesení poznatků a zkušeností do vzdělávání, a sice do předmětu digitalizace na Slezské univerzitě v Opavě. Podařilo se nám získat podporu Studentské grantové soutěže (SGS) v letech 2022–2024. V projektu SGS (Katuščák, 2024) jsme se zaměřili na české historické rukopisy psané kurentem.

Vycházeli jsme z hypotézy, že v Česku zatím není k dispozici dostatečně efektivní agregovaný model automatické transkripce, který by byl vytvořen na dostatečně vel-

5 Train set. Pomocí nástroje Transkribus Expert Client je možné cvičit (trénovat) model rozpoznávání rukopisného textu, aby bylo možné transkribovat automaticky sbírky dokumentů. Model je výsledkem cvičení, proto je při jeho tvorbě třeba cvičit tak, aby stroj rozpoznal určitý styl psaní v zobrazovaných obrázcích dokumentů a poskytl víceméně přesný přepis. Ke cvičení (TRAIN) modelu je zapotřebí 5 000 až 15 000 slov (přibližně 25–75 stran) přepsaného materiálu. Přepis se získá manuálním přepisem řádek po řádku přesně podle předlohy.

6 Ground Truth (GT. základní pravda) jsou přesné a ověřené údaje, které se používají pro trénování modelů strojního učení, jako jsou modely používané pro automatické přepisy v Transkribusu.

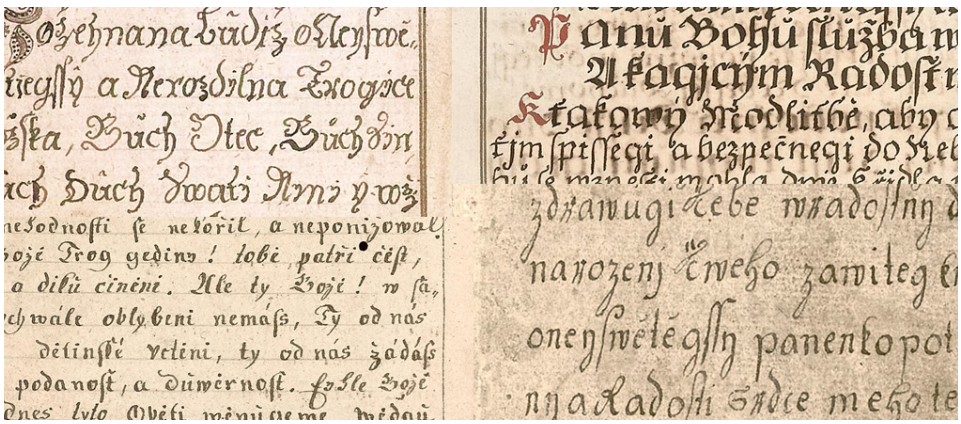
kém množství stran v kvalitě GT, jež by bylo možno použít pro tvorbu lepších modelů transkripce. Skvělou práci v tomto směru vykonává Anna Michalcová (2024). Důsledkem absence nástrojů automatické transkripce je, že historické dokumenty knihoven, muzeí, archivů a dalších institucí jsou sice digitalizovány, avšak jsou obvykle dostupné pouze jako digitální *faksimile*, *obrázky* (digitalizáty) bez transkripce. Tato vědecká úloha je spíše úkolem pro národní instituce než pro malé projekty typu SGS. Cílem daného malého projektu SGS bylo přispět k řešení problému transkripce a připravit odborníky, kteří postupně budou řešit důležitý úkol týkající se zpřístupnění historických dokumentů z českých a slovenských archivů, knihoven, muzeí apod.

V dalších kapitolách jsou stručně popsány aktivity studentů Oddělení knihovnictví Filozoficko-přírodovědecké fakulty Ústavu bohemistiky a knihovnictví Slezské univerzity v Opavě.

2 Tvorba modelů transkripce podle vybraných rukopisných náboženských dokumentů. Cesta k modelu transkripce Agreg-8 (Vojtěch Říha)

V oblasti automatické transkripce rukopisů došlo v nedávné době k významnému pokroku díky vytvoření funkčního studentského supermodelu, který dokáže transkribovat západoslovanské rukopisné texty z období 18.–20. století. Tento supermodel s názvem *CZECH supermodel_SGS* (ID 220865), fungující na bázi nástroje Transkribus, vznikl sloučením několika dílčích modelů, které vyvinuli studenti Slezské univerzity v Opavě.

Model Agreg-8 (207993), připravený studentem Vojtěchem Říhou, je trénovaný na rukopisném materiálu pěti různých sbírek, přičemž cílem bylo získat sadu podobných dokumentů, které by v konkrétních rysech přispěly k celistvosti modelu (časový rozsah 18. až 19. století, tematika modliteb a písní z katolického prostředí). Klíčovým aspektem však byla podobnost písma jednotlivých písařů, která zajistila kvalitní výsledek fungující v případě vybraného rukopisného stylu (viz obrázek 1).

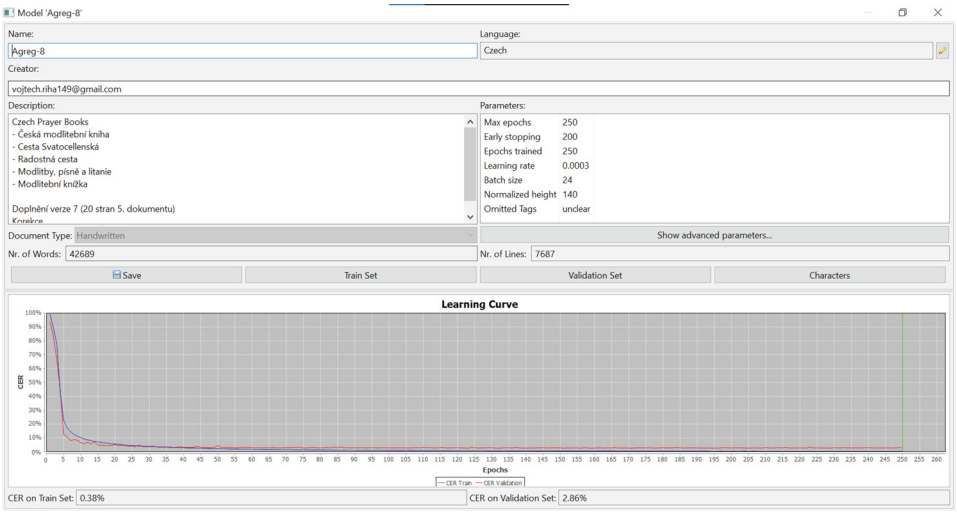


Obr. 1 Koláž s ukázkami rukopisných sbírek (vlastní koláž autora, části obrázků převzaty z databáze Manuscriptorium)

Vybrané sbírky jsou dostupné v digitální knihovně Manuscriptorium:

1. *Česká modlitební kniha* (ČESKÁ, 1733–1766)
2. *Cesta Svatocellenská* (CESTA, 1733–1766)
3. *Radostná cesta* (RADOSTNÁ, 1829–1884)
4. *Modlitby, písně a litanie* (MODLITBY, 1826)
5. *Modlitební knížka* (MODLITEBNÍ, 1700–1750)

Trénování modelu Agreg-8 probíhalo na 250 trénovacích cyklech 21 hodin a 37 minut. V procesu byl kladen důraz na kvalitní přepis, proto byla prováděna několikerá kontrola manuálně transkribovaného textu. Celkových 454 stran bylo poté označeno kvalitou *ground truth,* neboli „základní pravda", a tím zahrnuto do výsledného modelu. V trénovací sadě je obsaženo 42 842 slov, validační sada čítá 3 156 slov, souhrnně tak model dosahuje bez dvou slov počtu až 46 tisíc. Výsledná chybovost modelu (CER) se pohybuje okolo 0,4 % na trénovací sadě, na validační sadě pak 2,86 % (viz obrázek 2).



Obr. 2 Profil modelu Agreg-8 (archiv autorů)

Jinými slovy, při spuštění automatické transkripce se přepíše více než 97 znaků ze 100 znaků správně, což indikuje poměrně vysokou přesnost přepisu v případě konkrétních pěti zahrnutých sbírek. Je samozřejmé, že u jiných rukopisů nebude úspěšnost tak vysoká. Model Agreg-8 jsme nicméně experimentálně aplikovali na další vybraný rukopis nezahrnutý do žádného z datasetů a chybovost CER zůstala poměrně nízká (okolo 5 %). Průběh trénování lze sledovat na křivce učení, která ukazuje vstupní přesnost dat, neboť již po úvodních deseti epochách, jinými slovy trénovacích cyklech, model dosahoval chybovosti okolo 10 % a konzistentně se zlepšoval (viz obrázek 2).

Zaměřme se nyní na praktické možnosti při zpracovávání rukopisných dokumentů, neboť ve srovnání se starými tisky či strojopisem mají rukopisy zcela odlišnou genezi. Diference jednotlivých znaků je u ručně psaného dokumentu mnohem větší, tudíž je pro umělou inteligenci daleko obtížnější texty rukopisu rozpoznat. Mezi veřejnými modely

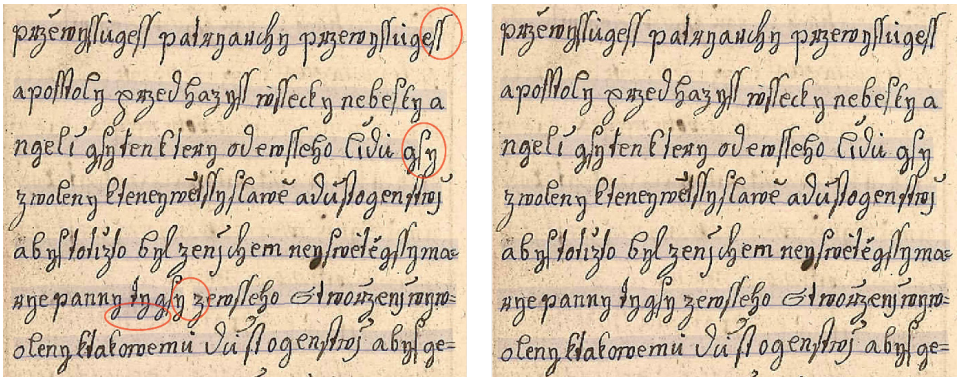
na platformě Transkribus zatím nemáme k dispozici řešení pro české bohemikální dokumenty vyjma projektu *Old Czech Handwriting* (Michalcová, 2022) a modelu *Moravian Land Records* (Schwarz, 2024). V současné době je tak pro transkripci vlastních dokumentů nutné trénování nového modelu, z širšího hlediska pak vyvstává potřeba vytvořit robustnější agregované modely nejen pro české rukopisy. Při transkripci je velmi důležitá samotná příprava, zvolení metod a přístupů, také však samotný proces, během kterého se výzkumník snaží aktivně reagovat na výsledky dílčích vlastních modelů a postupně je vylepšovat. Samotné trénování tak nespočívá pouze v přidávání dalších textových dat, byť tento postup znamená pro zdokonalení modelu značný přínos. Klíčovou činností se však stává celková optimalizace. Při tvorbě modelu HTR (*Handwritten Text Recognition*, rozpoznávání ručně psaného textu) je třeba dbát na několik zásadních aspektů. Výše zmíněný model Agreg-8 zde poslouží jako vzor.

Uvažme situaci, při níž jsme zvolili vhodný dokument k digitalizaci a zároveň zajistili kvalitní nasnímání všech stran. Operujeme zatím s obrázky, s množstvím pixelů, nikoli však s dokumentem, který obsahuje editovatelnou textovou složku, a proto je důležitá automatická transkripce. V současné době platforma Transkribus nenabízí velké množství modelů HTR, z tohoto důvodu se dále v textu věnujeme dílčím krokům, které vedou k tvorbě vlastního řešení.

Prvním krokem je tedy segmentace, jejíž pomocí dáváme formu celému obrázku, rozdělíme plochu na textové, případně obrázkové regiony a definujeme pozici textu, jinými slovy vytvoříme pro následnou automatizaci vzor pro rozpoznání jednotlivých řádků. V současné době existují již pokročilé segmentační modely, jež velkou část udělají automaticky v základu, nelze se však na tento postup nekriticky spoléhat. Kvůli nedokonalé segmentaci začínají vznikat první nepřesnosti, které se snažíme eliminovat vhodným nastavením. Není tedy zvykem provádět celou segmentaci manuálně, nýbrž postupným testováním nastavit parametry automatizace tak, aby fungovala pokud možno co nejpřesněji v rozsahu celého dokumentu. Případné chybné segmenty je pak nutné manuálně upravit. Možnosti nastavení automatické segmentace na dokumentu *Česká modlitební kniha* (ČESKÁ, 1733–1766) v prostředí platformy *Transkribus eXpert* na praktickém příkladě:

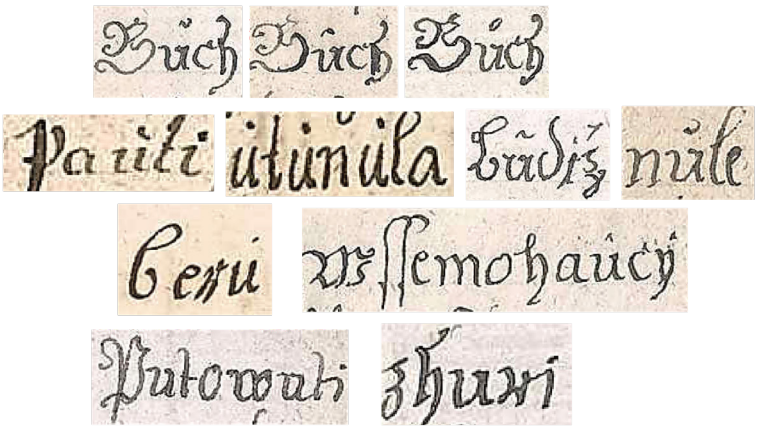
V prvé řadě vybíráme segmentační model. Je tedy důležité sledovat vlastnosti textu v celém dokumentu a dle toho zvolit variantu modelu. V případě nahodilého směru textových čar je vhodné aplikovat model tolerující směrové odchylky, jinak je tomu u dokumentů s rozložením homogenním, kde se vyplatí konzistentní segmentační modely. Běžnými problémy, se kterými se někdy setká prakticky každý, jsou nesprávně definované řádky, vynechané řádky, rozdělená slova apod.

V druhé řadě lze tedy experimentovat s nastavením základní čáry (tzv. *baseline*) a jejími parametry tak, abychom se vyhnuli nepřesnostem. Detekuje-li algoritmus jakýkoli rušivý element jako základní čáru, nabízí se zvýšit minimální délku této čáry tak, aby se na šum detekce nevztahovala. Řádek rozdělený do několika samostatných segmentů můžeme opravit snížením maximální vzdálenosti pro jejich slučování. Ačkoli se snažíme vstupní data importovat v co nejvyšší kvalitě, nemáme vždy k dispozici obrázky s požadovaným rozlišením, tudíž lze v nastavení pracovat také se škálováním obrázků. Na příkladu níže můžeme vidět výsledek segmentace před nastavením a snímek po optimalizaci parametrů (viz obrázek 3).



Obr. 3 Porovnání segmentace před a po úpravě parametrů (Manuskriptorium, upraveno)

Po úspěšně zvládnuté segmentaci se dostáváme k samotné transkripci, v našem případě ve tvorbě modelu pro daný dokument. Hlavním specifikem při práci s rukopisnými dokumenty je větší variabilita jednotlivých písmen. Různí písaři přirozeně zapisovali konkrétní znaky odlišnými způsoby, rukopisný styl se proměňoval i u jednoho autora během jeho života. Problém různosti zápisu znaků ve velké míře zasahuje také do každého dokumentu jednotlivě. Ať už vlivem nečitelnosti snímku nebo nepřesnosti pisatele velmi snadno nastane situace, ve které jsou si dvě různá písmena tak podobná, že by bez kontextu nebylo možné znaky od sebe rozeznat. Nejen v českých rukopisných dokumentech pak narážíme také na problém nečitelné diakritiky, která může být v jednom dokumentu psána mnohými způsoby (viz obrázek 4).



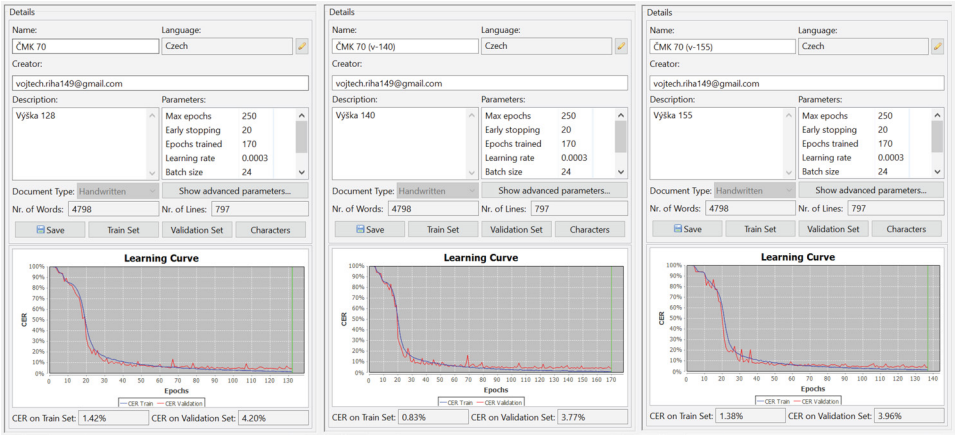
Obr. 4 Proměnlivost zápisu znaku „u“ (archiv autorů)

Dobová gramatika nebyla striktně vymezena jako dnes, jazyková norma nebyla dostatečně ustálena, jednotliví písaři si vytvářeli vlastní pravidla nebo psali zcela bez pravidel. Konkrétně psaní diakritických znamének pak nemuselo mít pouze funkci distinktivní v oblasti délky fonémů, nýbrž mohlo sloužit k rozlišení jednotlivých písmen od sebe navzájem. Na obrázku výše můžeme vidět písmeno „u“ v rozličných variantách (někde háček, kroužek, čárka, tečka, půlměsíc, vlnka, stříška apod., jinde dokonce diakritika chybí, ačkoli by dle dnešní

gramatiky z hlediska kvantity slovo vyžadovalo kroužek). Při transkripci takového textu je pak nutné vytvořit jednoduché pravidlo, podle kterého budeme při trénování postupovat, aby trénovaný model dokázal znak správně vyhodnotit. Jinými slovy, není potřeba hledat alternativu ke každé grafické anomálii, naopak je nezbytné snažit se jednotlivé znaky koordinovat a integrovat nejen podle jejich vzhledu, ale také podle jejich významu. Nemožnost dosažení 100% přesné transkripce u rukopisných dokumentů dokazuje také problematika zápisu písmen „i“ a „y“. Nejen že se dobová gramatika značně odlišovala od současné, ale také docházelo k netradičním zápisům znaků, které by nebyly u tištěných dokumentů možné (viz obrázek 5).

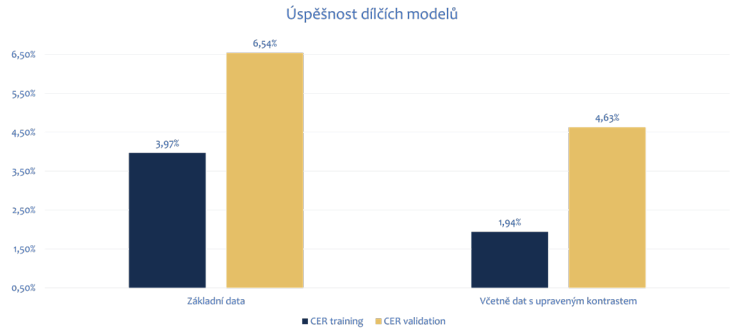
Obr. 5 Problematika podobnosti písmen „i“ a „y“ (archiv autorů)

Proces samotného trénování je přímo ovlivněný kvalitou manuálně přepsaných dat. Všechny strany zahrnuté jak do trénovací, tak do validační sady, by tedy měly být v kvalitě *ground truth*, jak již bylo řečeno výše. Zpětně pak při hodnocení dílčích modelů vycházíme z validační sady, jejíž chyby nám podávají informace o problémech modelu. Vedle další korekce a dalšího přidávání dat se však nabízí možnost experimentovat s parametry tréninku, jedním z nich je například výška řádku. Na příkladech níže (viz obrázek 6) lze sledovat dílčí model ČMK-70, u kterého jsme při tréninku s výškou řádku experimentovali. Konkrétně jsme zvýšili základní nastavení 128 pixelů na hodnotu 140 pixelů a poté až na 155 pixelů. Výsledky vykazují menší chybovost při nastavení parametru výšky řádku na číslo 140 pixelů, protože další zvyšování již začalo při transkripci operovat s vedlejšími řádky.



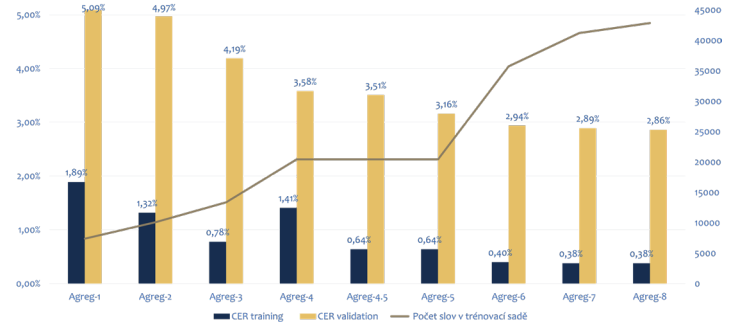
Obr. 6 Experiment s výškou řádku u modelu ČMK-70 (128px, 140px, 155px) (archiv autorů)

Některé rukopisné dokumenty mají na konkrétních místech v důsledku vybledlého inkoustu nebo poškozeného papíru těžko rozpoznatelná slova. Pro vyšší kvalitu modelu tak byla některá místa označena tagem „unclear“, aby se vynechaly nejednoznačnosti, které by učily model nepřesné transkripci. V trénovacím datasetu byly takto označené části z modelu vyřazeny. Vedle toho jsme provedli experiment s rozšířením obrazových dat *(image data augmentation)*. Jedná se o alternativní způsob navýšení rozmanitosti znaků, kdy se stávající obrazová data konkrétními technikami upraví, což má následně efekt dalšího snížení počtu chyb při automatické transkripci. Významnou augmentační technikou jsou fotometrické úpravy, tzn. změna kontrastu, jasu, ostrosti, šumu nebo úprava barevnosti apod. V našem experimentu jsme takto rozšířili méně obsáhlý dílčí model ČMK-20. Přidáním snímků s upraveným jasem a kontrastem bylo dosaženo významného snížení chybovosti, výsledky tak ukázaly, že je tato technika přinejmenším v začátcích trénování užitečným doplněním (viz obrázek 7).



Obr. 7 Augmentace obrazových dat modelu u ČMK-20 (archiv autorů)

Dalšími augmentačními technikami jsou geometrické transformace v podobě rotací, ořezů nebo škálování původních dat. Zajímavou a účinnou metodou rozšiřování je však také použití deformací, tzn. zvýšení či snížení výšky nebo délky konkrétních obrázků. Tento způsob augmentace dat byl použit v procesu trénování modelu Agreg-8, kam jsme zahrnuli snímky s upravenou výškou (změny na 80 %, 90 %, 110 % a 120 %). Výsledek tohoto experimentu lze sledovat na celkovém vývoji modelu Agreg, konkrétně porovnáním modelů Agreg-5 a Agreg-6, který již obsahuje data včetně augmentace (viz obrázek 8).



Obr. 8 Celkový vývoj úspěšnosti a počtu slov modelu Agreg (archiv autorů)

Na grafu výše (obr. 8), který ve sloupcové části ukazuje snižující se chybovost znaků a ve spojnicové části zvyšování počtu slov, lze také dokumentovat již popsanou změnu výšky řádků, se kterou jsme experimentovali mezi verzí Agreg-4 a Agreg-5. První verze Agreg-1 sdružovala dva pracovní modely „ČMK“ (data z *České modlitební knihy)* a „RC“ (data *z Radostné cesty*) a sloužila jako jádro celého tréninku. K tomuto modelu jsme postupně přidávali další data, další rukopisné sbírky s modlitební tematikou *(Cesta Svatocellenská* a *Modlitby, písně a litanie).* Mezi verzí modelu Agreg-6 až Agreg-8 lze poté sledovat přidání posledního rukopisného dokumentu Modlitební knížka.

3 Tvorba modelů transkripce na rukopisech J. H. A. Gallaše, F. Poláška a O. Jaroše (Lukáš Němec)

Cílem práce studenta Lukáše Němce bylo vytvořit na základě pečlivě vybraných rukopisů model, který by si uměl poradit s rukopisnými vzorky z 2. poloviny 18. století až do 1. poloviny století dvacátého, resp. dokázal by přečíst ručně psané texty západoslovanské provenience od dob národního obrození až do první republiky. Využití takového modelu nabízí široké spektrum možností. Archivy obsahují rozsáhlé množství textových materiálů z uvedeného období a použití tohoto modelu by mohlo výrazně usnadnit práci badatelům. Uplatnění lze nalézt například při analýze legionářských dopisů z období první světové války, válečných deníků vzniklých během bojů na východní frontě nebo při studiu rukopisných fragmentů a opomíjených textů méně známých či regionálních autorů českého národního obrození. Využití tohoto modelu je rovněž velmi přínosné při zpracování a analýze různých rodových, obecních či spolkových kronik, matrik, stejně jako historických katastrálních a pozemkových knih.

Pro tvorbu modelu byly z několika dalších možných rukopisných dokumentů selektivně vybrány tyto:

1. Gallaš, Josef Heřman Agapit [Rukopis]: *Mytické povídky o bozích a bohyních moravských Slovanů.* (Gallaš, 1820)
2. Gallaš, Josef Heřman Agapit: [Rukopis]. *Fyzické památky města Hranice a okolí.* (Gallaš, 1808–1811)
3. Gallaš, Josef Heřman Agapit [Rukopis] *Walaši v kraji Přerovském* (Gallaš, 1801–1804)
4. Polášek, František [Rukopis]: *Pravé poznání Boha aneb troje hodinky o dokonalostech božských [Rukopis]* (Polášek, 1800–1900)
5. Jaroš, Otakar [Rukopis]: *Nauka o terénu* [*Školní* sešit, čtverečkovaný/linkovaný papír]. (Jaroš)

Josef Heřman Agapit Gallaš, původem z Hranic, patří mezi tamější nejvýznamnější rodáky, byl vojenským polním lékařem a zakladatelem první hranické knihovny. František Polášek, katolický kněz, pocházel z městečka Příbor v okrese Nový Jičín. Otakar Jaroš, voják a válečný hrdina, patřil k významným studentům hranické vojenské akademie.

Při výběru vzorků jsme si stanovili několik podmínek s cílem zajistit co největší univerzálnost modelu, aby byl výběr autorů, rukopisů i témat co nejrozmanitější a nebyl striktně omezen pouze na jednu žánrovou oblast, například náboženství nebo literaturu.

Klíčovým bylo zaměřit se na autory spojené s naším regionem, protože jsme chtěli pracovat s texty, které jsou místně příslušné oblasti, kde žijeme nebo působíme a ke které máme citovou vazbu. Domníváme se, že tento záměr se nám podařilo naplnit beze zbytku, protože námi provedený výběr osobností je skutečně „multižánrový" a zároveň se vztahuje k našemu regionu.

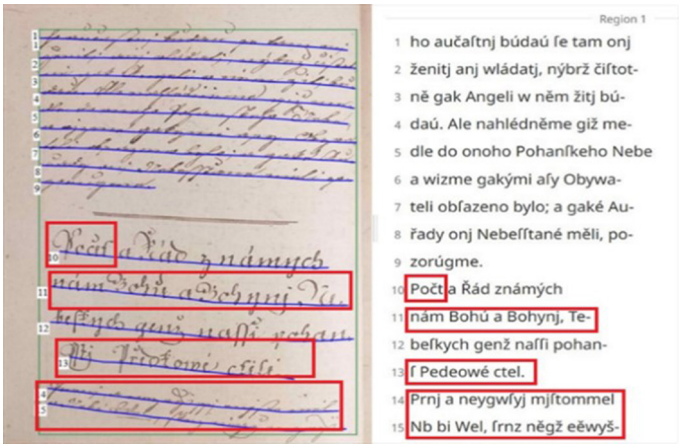
Práce na jednotlivých rukopisných dokumentech

Jedním z našich cílů bylo, aby alespoň některá z děl prošla celým procesem digitalizace, tj. od nasnímání digitalizátů přes vytvoření modelu, který by byl schopen textový obsah umístěný na nasnímaném obrázku přečíst, až po archivaci digitálních kopií v některém z repozitářů. To se povedlo u dvou Gallašových spisů (*Mytické povídky o bozích a bohyních* a *Fyzické památky města Hranice a okolí*) a obou rukopisů Otakara Jaroše. Uvedené rukopisy byly nasnímány zařízením ScanTent a pomocí aplikace DocScan nahrány do prostředí nástroje Transkribus, kde jsme s nimi dále pracovali na vytvoření modelu.

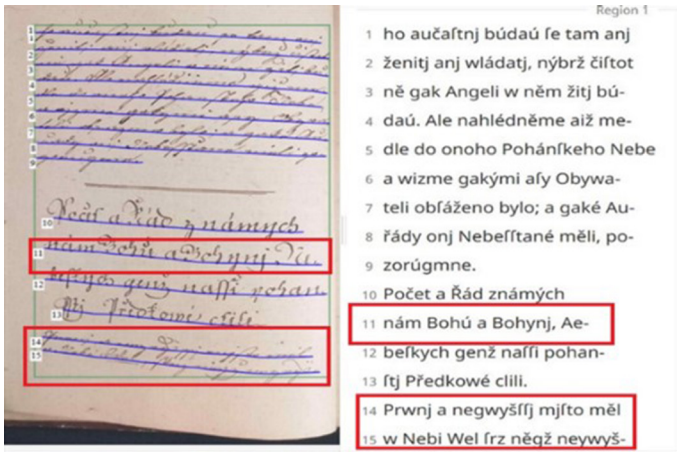
Obr. 9 Zařízení ScanTent, snímání rukopisu (archiv autora)

První rukopis, na kterém jsme začali pracovat, byly Gallašovy *Mytické povídky o bozích a bohyních* (viz obrázek 10). Rukopis se vyznačuje počínající degradací papíru a častým vypadáváním inkoustu. To vedlo k částečnému vyblednutí původního textu, který je navíc poměrně obtížně čitelný, což výrazně ztížilo proces osvojování si čtení dobového rukopisu. Gallašovy texty mají několik specifik, mezi něž patří časté gramatické chyby (obrázek 12), dále psaní podstatných jmen velkými písmeny, což byl zlozvyk pocházející patrně z němčiny, a nestálost v grafické realizaci některých fonémů, např. „š"; „á"; „ú"; „ž" a „g".

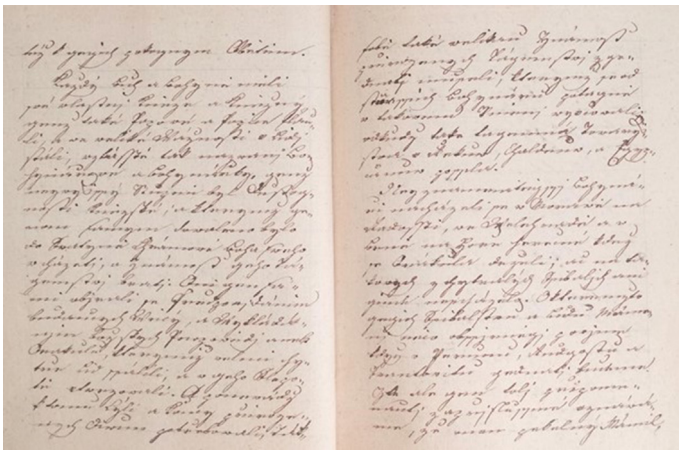
Model nazvaný *Mystic Absolut* (ID 210053) jsme vytvořili na základě 120 stránek GT s celkovým počtem kolem 23 tisíc slov na 4 470 řádcích a deklarovanou chybovostí 8,3 % na ověřovací sadě. Bohužel lepší výsledek i přes maximální snahu nebyl možný z důvodů uvedených výše. Teprve použití agregovaného modelu *Finale 2.0,* o němž se podrobněji zmíníme níže, vedlo ke snížení chybovosti při rozpoznávání textu. Tento výsledek byl způsoben schopností agregovaného modelu efektivněji zpracovávat i méně časté grafémy, k jejichž přesnější identifikaci přispěly dílčí modely, které tyto grafémy zahrnovaly (obrázek 10 a 11).



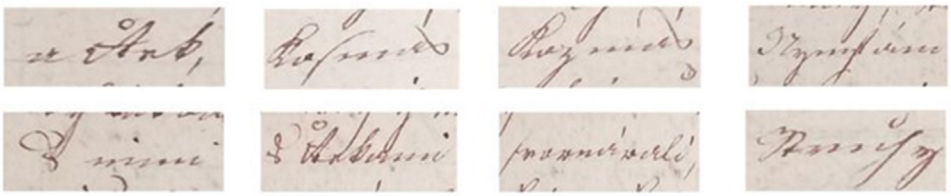
Obr. 10 Model Mystic Absolut s chybovostí 8,3 % (archiv autora)



Obr. 11 Agregovaný model Finale 2.0 s chybovostí 6,5 % (archiv autora)

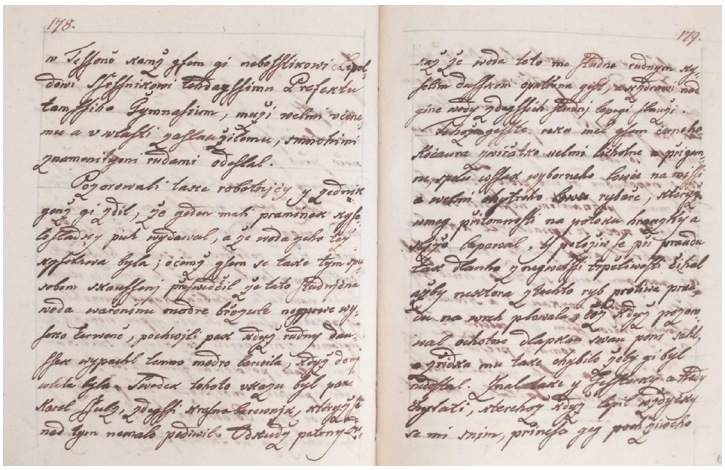


Obr. 12 Ukázka rukopisu Mytické povídky o bozích a bohyních (MZA Brno)



Obr. 13 Ukázka různých způsobů psaní grafémů a proměnlivého sklonu písma (archiv autora)

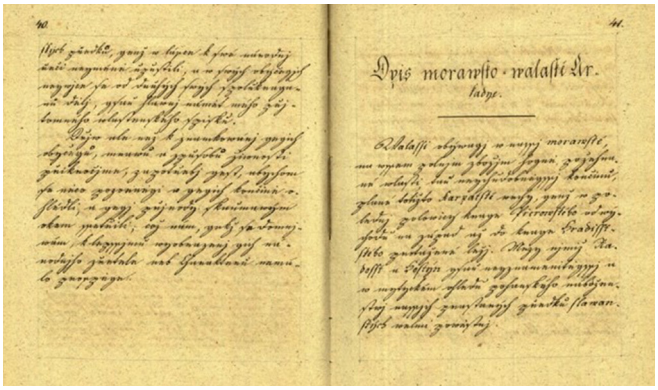
Druhý Gallašův rukopis *Fyzické památky města Hranice* (viz obr. 14) představoval úplně odlišný typ písma než vzorek první (obr. 13). Písmo bylo na první pohled úhlednější a čitelnější, avšak obsahovalo více písařských stylů, vzniklých patrně podle toho, jak unavená byla ruka. Zde jsme řešili prosvítání textu z protilehlých stránek a zasahování psaných grafémů do spodní části osnovy. Tento nešvar, kdy model někdy detekoval grafém jako další textovou linku, byl částečně odstraněn drobnou úpravou výšky řádku (ze 128 na 140 px), avšak stále je nutná částečná úprava textu, týká se zhruba 1 % všech možných znaků na stránce.



Obr. 14 Ukázka rukopisu Fyzické památky města Hranice a okolí (MZA Brno)

Na základě tohoto rukopisu, který obsahoval kolem 20 tisíc slov na 3 074 řádcích, byl vytvořen dílčí model *Physical Absolut* (ID 241489) s chybovostí 5,3 % v ověřovací sadě.

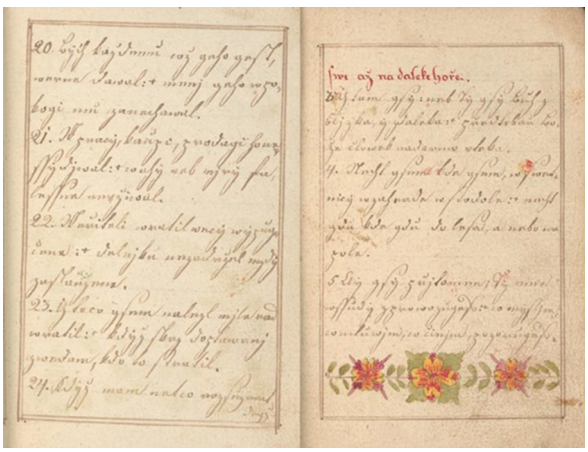
Jak je patrno z obr. 15, na třetím rukopisném vzorku z pera J. H. A. Gallaše *Walaši v kraji přerovském* byla pozoruhodná skutečnost, že ač je rukopis velmi odlišný od předešlého Gallašova rukopisu, model vytvořený na jeho základě pro tento typ písma velice dobře fungoval. Bohužel zde byly velké problémy s rozeznáváním diakritiky, vyžadující ruční korekci u grafémů „u“, „ů“, „ú“, „e“ a „ě“ (tento problém byl vyřešen agregovaným modelem Finale 2.0).



Obr. 15 Ukázka rukopisu Walaši v kraji přerovském (zdroj Manuskriptorium)

Na bázi uvedeného rukopisu byl po vložení 16 tisíc slov na 2 600 textových linkách vytvořen model *Walachian Absolut* (ID 211773) s chybovostí 5 %.

U rukopisu F. Poláška *Pravé poznání boha aneb troje hodinky o dokonalostech božských* (viz obr. 16) jsme při práci čelili výzvě spojené se třemi odlišnými druhy písma, přičemž dva z nich byly v textu zastoupeny jen v omezené míře. Tato skutečnost způsobovala občasné problémy s přesnou detekcí grafémů během procesu automatického rozpoznávání textu. Problém se však podařilo vyřešit za pomoci agregovaného modelu *Finale 2.0.* Agregovaný model totiž obsahoval jiné rukopisné vzorky s podobnými typy textu, což umožnilo lepší identifikaci méně častých grafémů a zvýšilo celkovou přesnost rozpoznávání.

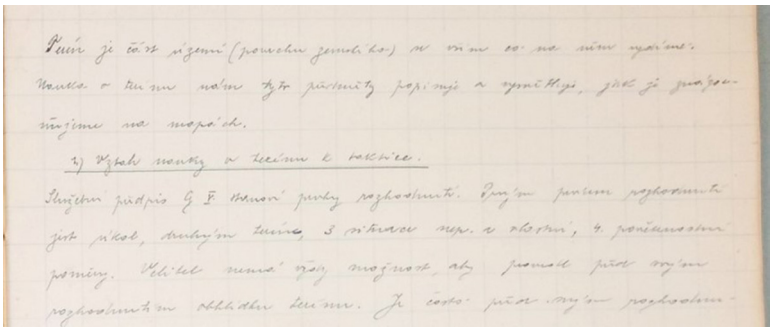


Obr. 16 Ukázka rukopisu Pravé poznání boha (zdroj Manuskriptorium)

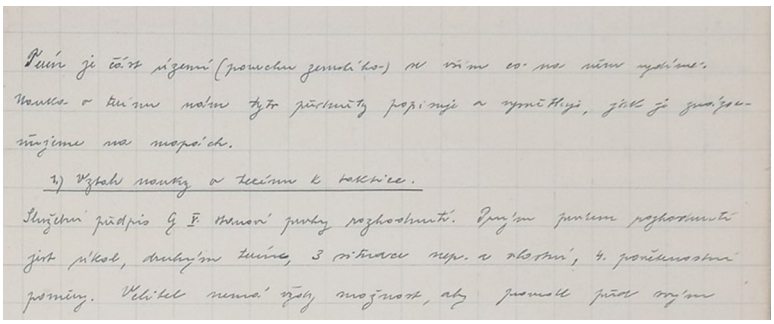
Tento dílčí model nazvaný *Franz II.* je ze všech dílčích modelů nejmenší (nejméně robustní), neboť k jeho vytvoření bylo použito pouze pěti tisíc slov, což je považováno za spodní hranici pro počet vložených slov. I přes tento „handicap" jeho chybovost činila solidních 7,5 %.

Rukopisy Otakara Jaroše představovaly skutečnou výzvu. Školní sešity s rukopisnými vzorky nesly známky degradace, pravděpodobně způsobené nevhodným skladováním

v prostorách vojenského muzea s vysokou vlhkostí. Tato degradace byla dále umocněna značným vyblednutím inkoustu na některých stránkách a skutečností, že text byl napsán na čtverečkovaný papír, což komplikovalo jeho čitelnost. Všechny výše uvedené skutečnosti představovaly při procesu snímání digitalizátů pomocí zařízení ScanTent závažný problém. Aplikace DocScan měla v automatickém režimu velké problémy se zaostřením snímku (aplikace vyfotila snímek, aniž by došlo k jeho kvalitnímu zaměření). To mohlo být způsobeno například tím, že optika fotoaparátu nedokázala dostatečně rychle zaměřit osvětlený předmět umístěný na tmavém pozadí. Proto bylo nutno přistoupit k použití ručního módu – snímek jsme zhotovili až po kontrole zaostření. Následně jsme provedli jemnou korekci snímku pomocí nástroje Zoner za účelem optimálního vyvážení kontrastu tak, aby bylo písmo zřetelné a dobře čitelné a zároveň linky čtverečkovaného papíru nebyly příliš zvýrazněné, což byl, jak se později ukázalo, další problém.



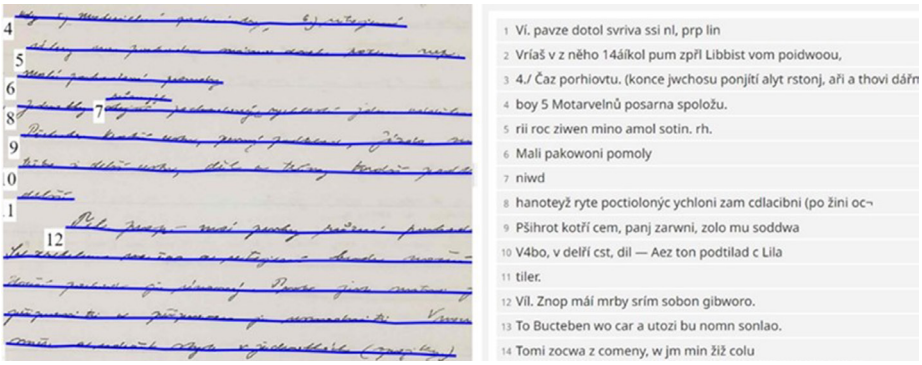
Obr. 17 Snímek pořízený v automatickém módu s patrnou neostrostí grafémů (archiv autora)



Obr. 18 Snímek upravený pomocí software Zoner (archiv autora)

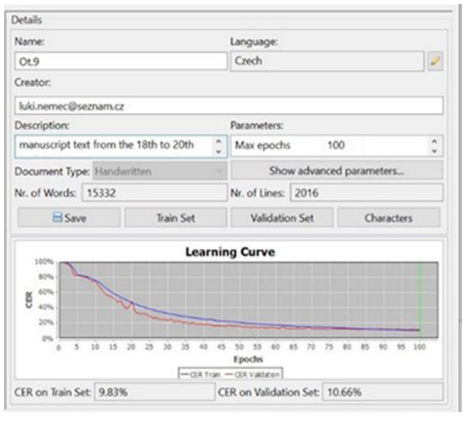
Před zahájením ručního vkládání textu jsme provedli experimentální testování dostupných modelů. Vybrali jsme model *Moravian Land Records* s udávanou chybovostí 6,4 %. Nepřesvědčivé výsledky schopností automatického rozpoznávání rukopisných znaků u uvedeného modelu (viz obr. 19) nás však přesvědčily o nutnosti vytvořit vlastní model pro dosažení vyšší přesnosti.

Práce na něm začala ruční segmentací stránky spolu s vkládáním textu, kdy jsme se potýkali se skutečností, že písmo, ač na první pohled úhledné, bylo špatně čitelné, a to zejména díky podobnosti grafémů „a“, „o“, „m“, „n“, „e“.

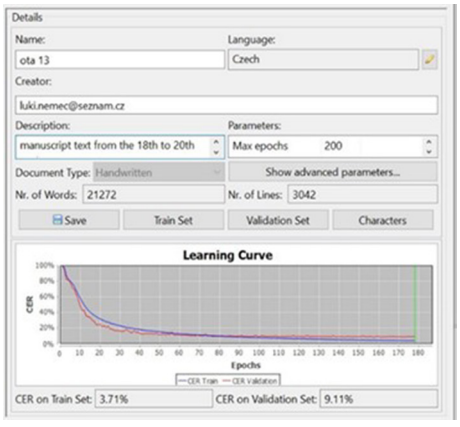


Obr. 19 Výsledek použití modelu Moravian Land Records s chybovostí 6,4 % (archiv autora)

Po ručním přepisu 15 332 slov na 2 016 základních linkách-řádcích jsme vytvořili první model, který vykazoval chybovost 10,88 % na ověřovací sadě, a to zejména u výše uvedených znaků.



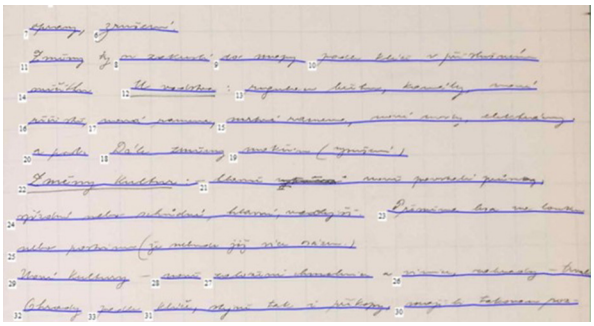
Obr. 20 Parametry prvního modelu (archiv autora)



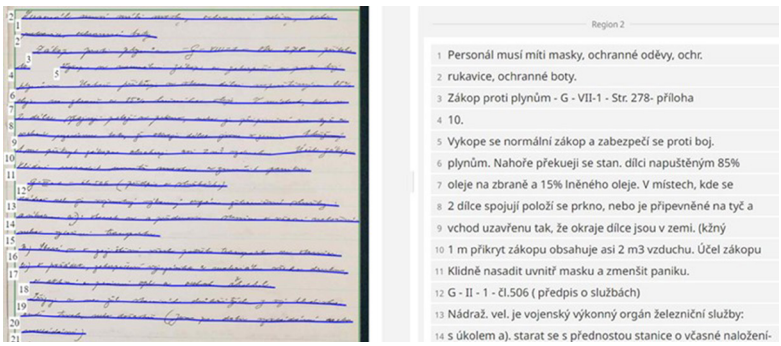
Obr. 21 Parametry dalšího modelu (archiv autora)

Z důvodu poměrně vysoké chybovosti (kolem 11 %) jsme pokračovali v ručním vkládání slov, aniž bychom vytvořený model použili při rekognici textu. Další model, nazvaný *Ota13,* vytvořený po vložení 21 272 slov na 3042 základních linkách, vykazoval chybovost 9,11 % na ověřovací sadě (viz obr. 21). Ten jsme již zkusili použít pro rozpoznávání textu na dalších stránkách rukopisu. Tento model však vykazoval určité nestandardní projevy, se kterými jsme se u jiných rukopisů nesetkali. Konkrétně docházelo k nesprávné segmentaci textu, při níž byly hlavní řádky rozdělovány na několik menších dílčích řádků (viz obr. 22).

Stejný model byl experimentálně vyzkoušen i na jiném rukopisu Otakara Jaroše, a to s diametrálně odlišným výsledkem, který nám dokázal funkčnost modelu a utvrdil nás v domněnce, že problémem bude pravděpodobně čtverečkovaný papír.



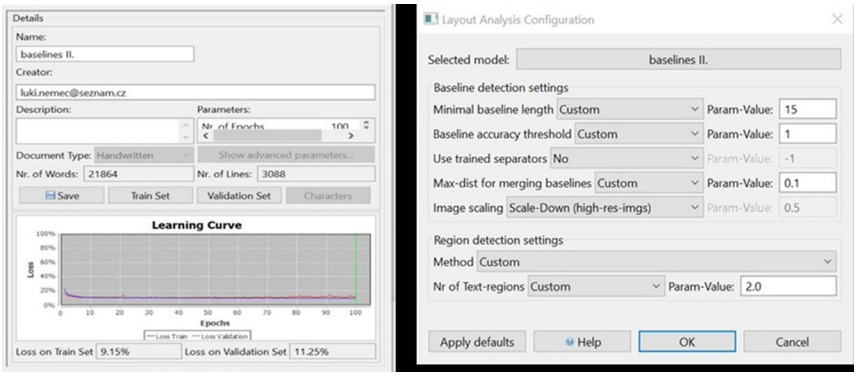
Obr. 22 Ukázka chybné segmentace modelu (archiv autora)



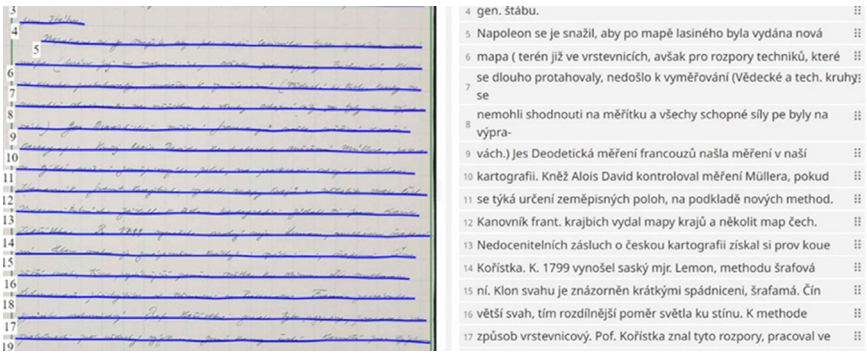
Obr. 23 Ukázka funkčnosti modelu na rukopisu stejného autora, avšak na jiném než čtverečkovaném papíru (archiv autora)

Jelikož jsme plánovali využít rukopis na čtverečkovaném papíru pro vytvoření dílčího modelu automatické transkripce jako součást budoucího agregovaného modelu *Finale 2.0,* rozhodli jsme se řešit problém chybné segmentace aplikací specializovaného modelu pro rekognici základních linek. Tento model, nazvaný *Basiclines II*, byl vytvořen na základě 3 088 ručně vložených základních linek s níže uvedenými parametry, přičemž jeho funkčnost byla následně ověřena (viz obr. 25).

I když chybovost na ověřovací sadě byla 11,25 %, segmentaci linek Jarošova rukopisu model *Basiclines II* prováděl bezchybně nebo pouze s minimem chyb, které byly odstraněny drobnou ruční korekcí. Proto následoval postup, kdy byl nejprve použit model *Basiclines II* určený pro segmentaci základních linek a pak provedena rekognice textu pomocí modelu Ota13 a vytrénování dalších modelů.



Obr. 24 Parametry modelu pro segmentaci základních linek (archiv autora)



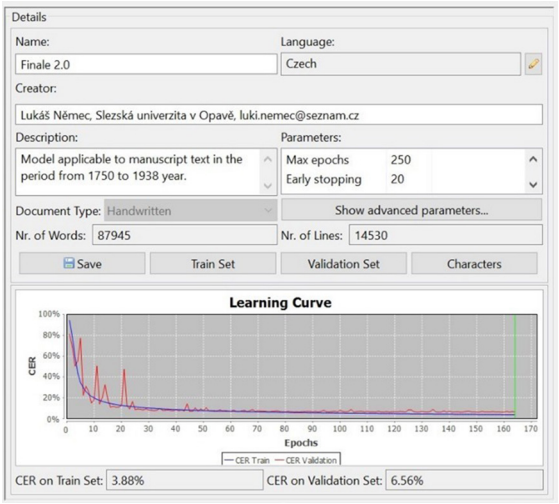
Obr. 25 Výsledek experimentu s využitím segmentace při rekognici stránky s použitím modelu *Ota13* (archiv autora)

Závěr našeho experimentu hovoří jednoznačně. V případě, že budeme provádět automatickou transkripci rukopisu psaného na podobně nestandardním podkladu a aplikovaný model bude vykazovat výše uvedené anomálie, doporučujeme zvážit následující postup:

1. S použitím většího množství již přepsaných stránek vytvořit model pro segmentaci základních linek, popř. využít stávající *Basiclines II.*
2. Upravit parametry pro segmentaci stránky s využitím parametrů uvedených na obrázku 24.
3. Správnost parametrů experimentálně ověřit na vybraných stránkách.
4. Aplikovat segmentační model na libovolné množství stran.
5. Pokračovat v transkripci textu metodou ručního vkládání textu nebo pomocí stávajícího modelu.

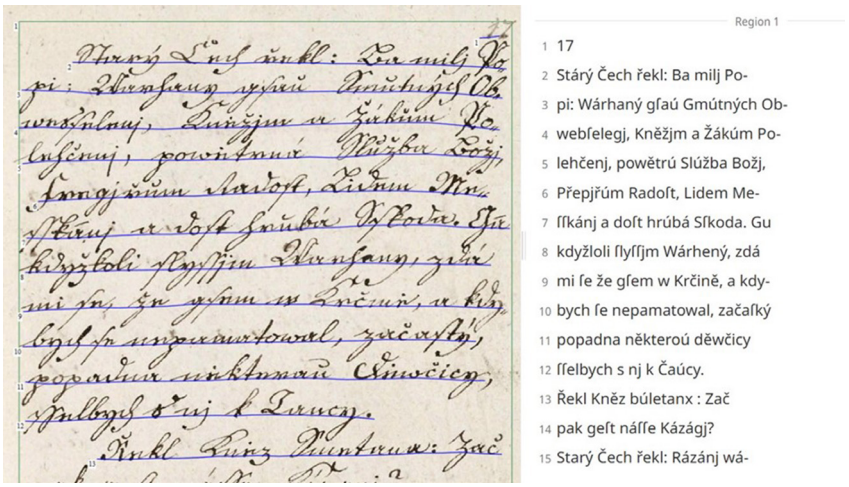
Konečný model Jarošova rukopisu byl nazván *Ota14* (ID 182965) a byl vytvořen na bázi 25 tisíc slov a 3 743 textových linek. Deklarovaná chybovost modelu je 7 % na ověřovací sadě. Vytvořením modelu *Ota14* jsme završili proces tvorby jednotlivých modelů. Po dokončení této fáze jsme přistoupili k integraci všech dílčích modelů do jednoho komplexního a univerzálního modelu, který jsme pojmenovali *Finale 2.0* (ID 213733). Tento agregovaný model představuje vyvrcholení naší práce, zahrnující veškeré získané

poznatky a optimalizace, které jsme aplikovali během vytváření jednotlivých modelů. Model *Finale 2.0* byl vytvořen tak, aby dosáhl pokud možno co nejvyšší přesnosti při zpracování širokého spektra rukopisných textů.



Obr. 26 Parametry modelu *Finale 2.0* (archiv autora)

Na závěr naší práce jsme provedli malý experiment zaměřený na ověření schopností modelu *Finale 2.0* při zpracování rukopisného textu, který byl náhodně vybrán z obsáhlého digitálního archivu Manuscriptorium. Vybraný rukopis nebyl do procesu vytváření modelu nijak zahrnut, což znamená, že model s tímto konkrétním písmem ani jeho charakteristickými znaky nemá žádnou předchozí zkušenost. Cílem bylo ověřit, jak si model poradí s neznámým materiálem a do jaké míry je schopen generalizovat své schopnosti při čtení textů, které nejsou součástí jeho tréninkového datasetu (viz obr. 27). Vybrali jsme rukopis *Rozličné písně starožitné* (1799).



Obr. 27 Ukázka schopnosti modelu *Finale 2.0* na jemu neznámém rukopisném vzorku (archiv autora)

Námi představený model tvoří nedílnou součást širšího a komplexnějšího agregovaného modelu s názvem *CZECH supermodel_SGS*. Tento agregovaný model dosahuje chybovosti pouhých 5,8 % na ověřovací sadě. Jeho vývoj byl realizován ve spolupráci s kolegy ze Slezské univerzity v Opavě. Domníváme se, že dosažené výsledky představují významný krok kupředu v oblasti automatického rozpoznávání rukopisných textů západoslovanského původu. Naše práce tak přispívá nejen ke zlepšení technologického zpracování historických textů, ale také k rozšíření možností jejich vědeckého zkoumání.

4 Transkripce strojopisných dokumentů (Klára Pohlová)

Pro bohemikální strojopisné dokumenty ještě stále neexistuje spolehlivý nástroj, který by byl schopný vykonat jejich automatickou transkripci. O to více se zde tedy poukazuje na potřebu vytvoření specifického modelu, použitelného pro strojopisné dokumenty. Cílem experimentu bylo ověřit existující veřejně dostupné modely transkripce strojopisu, případně vytvořit nový model transkripce strojopisných dokumentů vhodný pro budoucí použití.

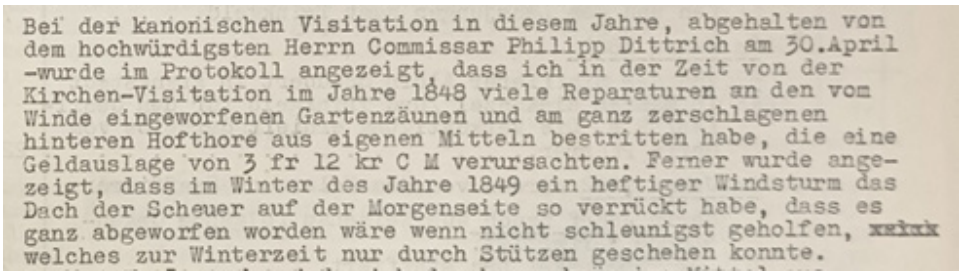
Ověření existujících veřejných modelů transkripce strojopisu

Databáze platformy Transkribus nabízí veřejné modely pro práci se strojopisnými dokumenty, většina z nich však nepracuje s češtinou, její gramatikou a interpunkčními znaménky. Ve výsledku je tedy text špatně rozpoznán a chybovost pak příliš velká, než abychom mohli přepis vyhlásit za úspěšný.

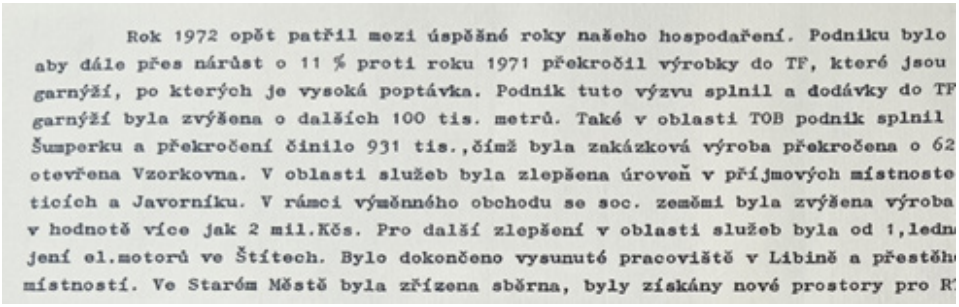
Vytvoření základní modelové báze vzorku strojopisných dokumentů s různými typy písma

Prvním krokem bylo hledání (heuristika) vhodných tiskopisů, které by byly vhodné pro pozdější použití při tvoření nového obecného použitelného modelu. Cílem tedy bylo najít zhruba 10 různých strojopisů s různými typy/fonty písma. Za tímto účelem byl vybrán Státní okresní archiv Jeseník, kde tiskopisy tvoří více jak polovinu veškerého archivního fondu.

Ukázky písma ve vybraných dokumentech (obr. 28 a 29):


Bei der kanonischen Visitation in diesem Jahre, abgehalten von
dem hochwürdigsten Herrn Commissar Philipp Dittrich am 30.April
-wurde im Protokoll angezeigt, dass ich in der Zeit von der
Kirchen-Visitation im Jahre 1848 viele Reparaturen an den vom
Winde eingeworfenen Gartenzäunen und am ganz zerschlagenen
hinteren Hofthore aus eigenen Mitteln bestritten habe, die eine
Geldauslage von 3 fr 12 kr C M verursachten. Ferner wurde ange-
zeigt, dass im Winter des Jahre 1849 ein heftiger Windsturm das
Dach der Scheuer auf der Morgenseite so verrückt habe, dass es
ganz abgeworfen worden wäre wenn nicht schleunigst geholfen, xxxxx
welches zur Winterzeit nur durch Stützen geschehen konnte.

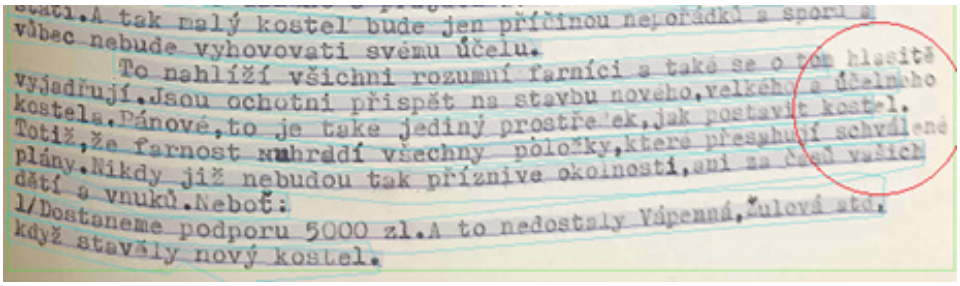
Obr. 28 Ukázka písma dokumentu NAD 197 (archiv autora)


Rok 1972 opět patřil mezi úspěšné roky našeho hospodaření. Podniku bylo
aby dále přes nárůst o 11 % proti roku 1971 překročil výrobky do TF, které jsou
garnýží, po kterých je vysoká poptávka. Podnik tuto výzvu splnil a dodávky do TF
garnýží byla zvýšena o dalších 100 tis. metrů. Také v oblasti TOB podnik splnil
Šumperku a překročení činilo 931 tis.,čímž byla zakázková výroba překročena o 62
otevřena Vzorkovna. V oblasti služeb byla zlepšena úroveň v příjmových místnoste
ticích a Javorníku. V rámci výměnného obchodu se soc. zeměmi byla zvýšena výroba
v hodnotě více jak 2 mil.Kčs. Pro další zlepšení v oblasti služeb byla od 1,ledn
jení el.motorů ve Štítech. Bylo dokončeno vysunuté pracoviště v Libině a přestěh
místností. Ve Staróm Městě byla zřízena sběrna, byly získány nové prostory pro R

Obr. 29 Ukázka písma dokumentu NAD 595 (archiv autora)

Práce v platformě Transkribus

Skeny textů byly nahrány do platformy Transkribus. Nejprve byla provedena segmentace řádků, kdy každý segmentovaný řádek se přesně dle originálu přepisuje pod segmentovanou část textu. Přepsáno bylo cca 4–6 stran textu. Dále proběhla kontrola segmentace u všech nahraných dokumentů. Ve většině případů byla segmentace správná, upravit ji bylo potřeba u minima stran. Nejvíce „náchylné“ ke špatné segmentaci byly strany, které měly text po stranách lehce rozmazaný, a tedy hůře čitelný. Bylo potřebné segmentovaný řádek prodloužit o neoznačený text.


Stačí.A tak malý kostel bude jen příčinou nepořádků a sporů a
vůbec nebude vyhovovati svému účelu.
To nahlíží všichni rozumní farníci a také se o tom hlásitě
vyjadřují.Jsou ochotni přispět na stavbu nového,velkého a účelného
kostela.Pánové,to je také jediný prostředek,jak postavíte kostel.
Totiž,že farnost zahrdí všechny položky,které přesahují schválené
plány.Nikdy již nebudou tak příznivé okolnosti,ani za časů vašich
dětí a vnuků.Neboť:
1/Dostaneme podporu 5000 zl.A to nedostaly Vápenná,Žulová atd.
když stavěly nový kostel.

Obr. 30 Špatně provedená segmentace u konce řádků textu (archiv autora)

Dalším krokem bylo zahájení přepisu metodou HTR, použit byl model *Stroj1*, který byl vytvořen na základě manuálně přepsaných stran, které již měly status GT, tedy *Ground Truth*. Tento krok je důležitý, pokud chceme docílit vytvoření vlastního modelu transkripce. Každá přepsaná strana prošla důkladnou kontrolou, chyby a rozdíly se manuálně přepsaly do správného tvaru dle originálu.

Ověření dostupných modelů

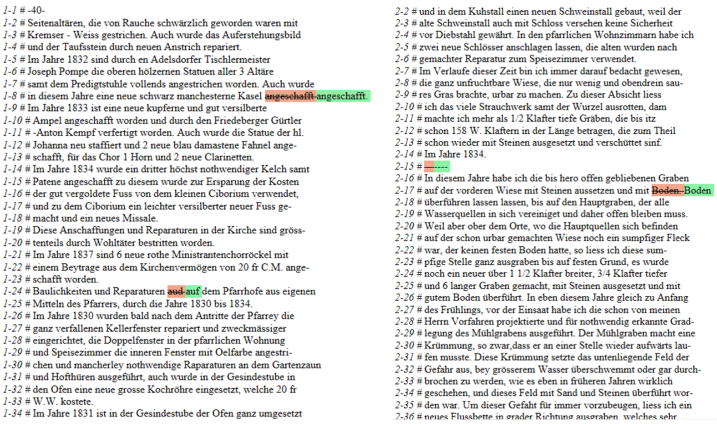
Byla vyzkoušena i možnost použít již existující modely strojopisných dokumentů, které má Transkribus ve své paměti. V případě úspěšnosti by nebylo potřeba tvořit nový model. Bohužel tyto pokusy úspěšné nebyly, a to vzhledem k tomu, že žádný z modelů neumí pracovat s českou diakritikou. Např. model 56926, byť je určen pro dokumenty psané psacím strojem, měl s textem v českém jazyce značný problém a výsledky jsou naprosto nepoužitelné:

Obr. 31 Text přepsaný s použitím modelu 56926 (archiv autora)

Tvorba vlastního modelu transkripce strojopisu

Z výše uvedených výsledků je zřejmé, že žádný z již dostupných modelů není ideálně použitelný pro texty v českém jazyce a je potřebné vytvořit vlastní model. Pro vytvoření nového modelu bylo rozhodnuto využít modely a data již existující a ověřené. Na jejich základě byl vytvořen a vytrénován model s označením 58379 *Slovak and Czech Typewriting*, který je nyní uložen v systému Transkribus jako veřejný model. Aktuálně je jeho chybovost (CER) 4,10 %, což je určitě výsledek, s nímž lze pracovat.

Nový model byl otestován na dvou vybraných dokumentech, jak v českém, tak německém jazyce.



Obr. 32 Přepisy nově vytvořeným modelem 58379 (archiv autora)

Výsledky jsou nadmíru úspěšné. Chybovost, CER, vychází 0,08. Text je tedy přepsán na 99,92 % správně (např. u slova „*angeschafft/angeschafft.*“ chybí tečka, slovo „*aud* bylo nahrazeno */auf*.

Jako druhá se zkoušela první dvojstrana dokumentu psaného v českém jazyce, NAD 214, viz obr. 33:

Obr. 33 Přepisy vytvořeným modelem 58379 s nulovou chybovostí (archiv autora)

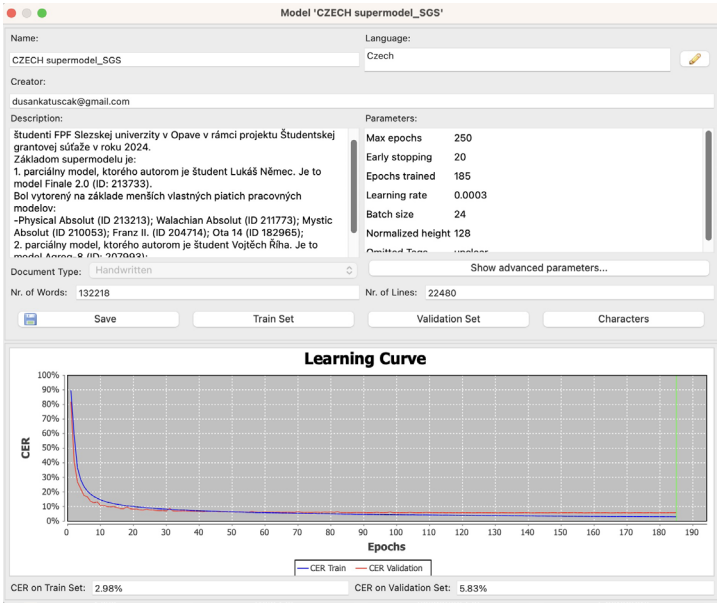
Zde je chybovost transkripce relativně nového strojopisu v podstatě nulová, text byl přepsán bez chyby, CER je tedy 0,00 %.

5 Závěr

Výzkumníci v projektu Studentské grantové soutěže (SGS) zvládli v průběhu několika měsíců práci v platformě Transkribus. Osvojili si metody přípravy, nahrávání, segmentace a provedli množství dílčích experimentů při tvorbě vlastních modelů transkripce. Získali znalosti, dovednosti a cenné know-how v transkripci rukopisů. Lukáš Němec vytvořil na základě pěti rukopisných dokumentů model *Finale 2.0* (ID: 213733) s chybovostí CER jen 6,56 %. Vynikající práci odvedl rovněž Vojtěch Říha, když jeho model model *Agreg-8* (ID: 207993) vykazoval chybovost pouhých 2,86 %. Do studie jsme také zařadili také popis a výsledky přípravy modelu ID 58379 pro transkripci strojopisných dokumentů Kláry Pohlové (Pohlová, 2024) z projektu SGS v roce 2023 (s chybovostí jen 4,10 %). Její parciální model ID 58379 byl pak zahrnut do supermodelu ID78289 (SUPERMODELP&T1, 2024).

V projektu SGS 2024 jsme nakonec vytvořili agregovaný *CZECH supermodel_SGS* (ID 220865) na základě výše uvedených parciálních modelů, které připravili studenti Lukáš Němec a Vojtěch Říha, a to s chybovostí jen 5,86 %. S naším modelem lze transkribovat podobné rukopisy s přesností 94,17 %. Základem supermodelu *CZECH supermodel_SGS* je:

1. parciální model, jehož autorem je student Lukáš Němec. Jedná se o model *Finale 2.0* (ID: 213733). Byl vytořen na základě menších vlastních pěti pracovních modelů: *Physical Absolut* (ID 213213); *Walachian Absolut* (ID 211773); *Mystic Absolut* (ID 210053); *Franz II.* (ID 204714); *Ota 14* (ID 182965);
2. parciální model, jehož autorem je student Vojtěch Říha. Jedná se o model *Agreg-8* (ID: 207993);
3. 15 rukopisných stran v kvalitě Ground Thruth (GT) z dokumentu Protokoly Matice slezské (ID:1663382).



Obr. 34 Czech supermodel SGS (archiv autorů)

Literatura

CESTA, 1733–1766. *Cesta Svatocellenská*. Online. 1733–1766. Josef Jan HÁJEK (písař). In: Jindřichův Hradec: Muzeum Jindřichohradecka, RK 037. Dostupné z: *https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CAIPDIG-MJ____RK_037_____3GYFBF3-cs?lang=cs. [cit. 2025-04-03].*

ČESKÁ, 1733–1766. *Česká modlitební kniha*. Online. 1733–1766. In: Jindřichův Hradec: Muzeum Jindřichohradecka, RK 071. Dostupné z: *https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CAIPDIG-MJ____RK_071_____176AYW2-cs?lang=cs. [cit. 2025-04-03].*

GALLAŠ, Josef Heřman Agapit, 1801–1804. *Walaši v kraji Přerovském*. Online. 1801–1804. Dostupné z: *https://new.manuscriptorium.com/apis/resolver-api/cs/catalog/default/detail/manuscriptorium%7CKNM___NMP___II_F_12_____1PU4RI1-cs*. [cit. 2025-04-03].

GALLAŠ, Josef Heřman Agapit, 1808–1811. [Fyzické]. Památky města Hranice a okolí [Rukopis]. In: *MZA Brno, G 11 Sbírka rukopisů Františkova muzea Brno, sign. 658, čeština, latina, papír, rukopisná kniha, originál, vázáno v tvrdých deskách, šířka 215 mm, výška 270 mm, pův. pag. 236, nová fol. 128; stará sign.: Schr. 223, pův. 288, červ.* 1808–1811. Podle L. Scholze (2006) je Gallašův rukopis Památek dnes uložen v Moravském zemském archivu v Brně (fond E6, kart. 490, sign. Oa7–12), je rozdělen na tři „epochy" (tj. díly), z nichž epocha třetí se dělí na čtyři samostatné svazky. In. Libor Scholz, Památk.

GALLAŠ, Josef Heřman Agapit, 1820. Mytické povídky o bozích a bohyních moravských Slovanů [Rukopis]. In: *MZA Brno, G 11, sign. 838, čeština, papír, rukopisná kniha, originál, vázáno v tvrdých, polokožených deskách, šířka 195 mm, výška 250 mm, stopy po pův. pag., starší fol. 125; stará sign.: Schr. 224, pův. 287, červ.* 1820.

GIANT, 2023. *The German Giant I*. Online. 20. March 2023. Dostupné z: https://app.transkribus.org/models/text/50870. [cit. 2025-04-03].

HRADIŠ, Michal et al, 2024. *DCGM / pero-ocr*. Online. 16. 12. 2024. Dostupné z: *https://github.com/DCGM/pero-ocr*. [cit. 2025-04-03].

JAROŠ, Otakar. Nauka o terénu, školní sešit, čtverečkovaný papír/linkovaný papír [školní sešit] Uloženo v: Historická expozice 71. mech. praporu „Sibiřského" v Hranicích. Zapůjčeno z pozůstalostní sbírky rodiny. In: *Uloženo v: Historická expozice 71. mech. praporu „Sibiřského" v Hranicích. Zapůjčeno z pozůstalostní sbírky rodiny.* Digitalizát vytvořen s laskavým svolením kurátora muzea nrtm. Radima Cába.

KACH, 1667. *Kach und Einmachbuch von Allerley Eingemachten Sachen von Zucker, Hänig und al/er Friichten, auch und erschiedlicher gueten Speisen* [Rukopis]. Online. 1667. Projekt SGS Slezská univerzita v Opavě. Dostupné z: *https://beta.transkribus.eu/collection/114429/doc/1154832/detail/6?view=combined&key=CDKKPGCBSBLBOOPSZHXFRUMI*. [cit. 2025-04-03].

KATUŠČÁK, Dušan, 2020a. *Digital humanities a automatická transkripcia rukopisných textov*. Online. Dostupné z: *https://itlib.cvtisr.sk/clanky/clanek3698/.* [cit. 2025-04-03].

KATUŠČÁK, Dušan, 2020b. *Najnovšie poznatky z výskumu automatického rozpoznávania textov historických dokumentov*. In: *Sborník z konference konané ve dnech 11.–13. 2. 2020*. Online. Dostupné z: *http://k21.fpf.slu.cz/wp-content/uploads/2020/12/Sbornik_K21_2020_RC.pdf*. [cit. 2025-04-03].

KATUŠČÁK, Dušan, 2022a. *Metodológia a metodika transkripcie historických textov*. Online. Projekt APVV UMB Skriptor. ISBN 978-80-557-2020-3. Dostupné z: *http://dx.doi.org/10.24040/2022.9788055720203*. [cit. 2025-04-03].

KATUŠČÁK, Dušan, 2022b. *Umelá inteligencia pomáha sprístupňovať písomné dedičstvo*. Online. *Knihovna: knihovnická revue*. Roč. 33, č. 2. ISSN 1802-8772. Dostupné z: *https://knihovnarevue.nkp.cz/archiv/2022-2/recenzovane-prispevky/umela-inteligencia-pomaha-spristupnovat-pisomne-dedicstvo*. [cit. 2025-04-03].

KATUŠČÁK, Dušan, 2022c. *Výkladový slovník pojmov a termínov* [platforma Transkribus]. Online. ISBN 978-80-557-2020-3. Dostupné z: *https://dx.doi.org/10.24040/2022.9788055720203*. [cit. 2025-04-03].

KATUŠČÁK, Dušan, 2024. *Vytvoření modelu pro automatický přepis českých historických rukopisů: od digitální faksimile k editovanému textu. [Projektová žádost SGS SU Opava]*. Opava: Slezská univerzita, s. 3, Projektová žádost Studentská grantová soutěž, Slezská univerzita.

KATUŠČÁK, Dušan a NAGY, Imrich, 2020. *Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom automatickej transkripcie historických rukopisov.* Katedra histórie, Univerzita Mateja Bela. Banská Bystrica: Univerzita Mateja Bela a Štátna vedecká knižnica. Agentúra na podporu vedy a výskumu SR; Vedecký projekt aplikovaného výskumu; Projekt podporený 170 000 Eur. APVV-19-NEWPROKECTZ-17816.

KATUŠČÁK, Dušan a NAGY, Imrich et al., 2023. *Automatická transkripcia historických dokumentov: metodická príručka na prácu s platformou Transkribus*. Online. 1. vyd. Banská Bystrica: Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici. ISBN 978-80-557-2070-8. Dostupné z: *https://doi.org/10.24040/2023.9788055720708*. [cit. 2025-04-03].

KATUŠČÁK, Dušan a NAGY, Imrich et al., 2024. *Automatická transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus: metodická príručka pre účastníkov workshopu*. Online. ISBN 978-80-557-2143-9. Dostupné z: *https://dx.doi.org/10.24040/2024.9788055721439.* [cit. 2025-04-03].

MICHALCOVÁ, Anna, 2022. *Padeřovská bible. Old Czech Handwriting [dataset].* Anna Michalcová s kolektivem. ID modelu: 58856.

MICHALCOVÁ, Anna et al., 2024. *HTR Winter School 2023/2024 – Medieval Czech – New Testament of Martin Lupáč (ONB Cod. 3304)* [dataset]. Online. 5. 2. 2024. Výber polodiplomaticky prepísaných textov z Nového zákona Martina Lupáča (1440, 320 × 210 mm, staročeština). Texty prepísali účastníci Zimnej školy HTR 2023/2024 vo Viedni. Dostupné z: *https://zenodo.org/records/10619017*. [cit. 2025-04-03].

MODLITBY, 1826. *Modlitby, písně a litanie*. Online. 1826. František PICHLER (písař). In: Brno: Moravské zemské muzeum, ST 2193. Dostupné z: *https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CAIPDIG-MZM___ST_2193_____2VNLSIA-cs?lang=cs*. [cit. 2025-04-03].

MODLITEBNÍ, 1700–1750. *Modlitební knížka*. Online. 1700–1750. In: Jindřichův Hradec: Muzeum Jindřichohradecka, RK 087. Dostupné z: *https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CAIPDIG-MJ____RK_087_____0RCYAF8-cs?lang=cs*. [cit. 2025-04-03].

NAGY, Imrich, 2024. Model ID: 197573. [transkribus expert]. *SKRIPTOR_Acta comitatus Nitriensis sedis iudiciariae final.*

NOCKELS, Joseph, GOODING, Paul a TERRAS, Melissa, 2024. Are Digital Humanities platforms facilitating sufficient diversity in research? A study of the Transkribus Scholarship Programme. In: *Digital Scholarship in the Humanities.* Online. 16. 04. 2024. ISSN 2055-768X. Dostupné z: *https://doi.org/10.1093/llc/fqae018*. [cit. 2025-04-03].

POHLOVÁ, Klára, 2024. *Automatická transkripce strojopisných dokumentů psaných v českém jazyce*. Online, diplomová práce. Dostupné z: *https://theses.cz/id/upmh25/?lang=sk*. [cit. 2025-04-03].

POLÁŠEK, František, 1800–1900. *Pravé poznání Boha aneb troje hodinky o dokonalostech božských* [rukopis]. Manuscriptorium. Online. [Datum: 15. 12. 2024]. Dostupné z: *https://new.manuscriptorium.com/apis/resolver-api/cs/catalog/default/detail/manuscriptorium%7CVMO___-VMO___K_24073_____0U6ABL2-cs*. [cit. 2025-04-03].

RADOSTNÁ, 1829–1884. *Radostná cesta*. Online. František PICHLER (písař). In: Brno: Moravské zemské muzeum, ST 2272. Dostupné z: *https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CAIPDIG-MZM___ST_2272_____1CKG86B-cs?lang=cs*. [cit. 2025-04-03].

READ, 2024. *READ COOP Transkribus*. Online. Dostupné z: *https://readcoop.eu/*. [cit. 2025-04-03].

ROZLIČNÉ, 1799. *Rozličné písně starožitné*. In: Moravská zemská knihovna v Brně pod signaturou RKP-0048.022. Dostupné z: *https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CMZK -MZKB RKP_0048_0222RSPJ43-xx?lang=cs*. [cit. 2025-04-03].

SCHWARZ, Johannes Georg, 2024. Moravian Land Records [Dataset]. In: *Dokumenty pocházejí převážně z MZA Brno and SOkA Znojmo.* CER Training 5,40 %, CER Validation 6,40 %. Tento model je jak pro češtinu, tak pro němčinu. ID modelu: 66429.

SMIDA, Matej, 2023. *Možnosti automatickej transkripcie v platforme Transkribus na príklade správ o vybavovaní sťažností občanov v období komunistickej diktatúry*. Online, diplomová práce. ISSN 1336-9148 a ISSN 2453-7845. Dostupné z: *https://doi.org/10.24040/ahn.2023.26.01.125-148*. [cit. 2025-04-03].

SUPERMODEL_M1, 2024. Slovak Supermodel M1 (SSM1) [Dataset]. *Zenodo.* Online. 1. ver., 24. 4. 2024. Jazyky zdrojových dokumentov: Slovak, Latin, Hungarian, Czech. Autori použitých datasetov: Katuščák, D., Nagy, I., Maliniak, P., Kurhajcová, A., Tomeček, O., Kunec, P., & Bôbová, M. ID modelu Transkribus: ID63569. Dostupné z: *https://doi.org/10.5281/zenodo.11109087*. [cit. 2025-04-03].

SUPERMODELP&T1, 2024. Slovak Supermodel P&T1 (SSPT1). Datasety projektu SKRIPTOR Univerzity Mateja Bela (First version (20240520)) [dataset]. Online. 20. 5. 2024. Autori parciálnych datasetov GT: Katuščák, D., Nižníková, L., Mikušková, M., Halfarová, N., Gajdošová, T., Málková, L., Taufrová, N., Nagy, I., Kováčová-Pohlová, K., Šmida, M., & Kociánová, N. ID modelu Transkribus: ID78289. Dostupné z: *https://doi.org/10.5281/zenodo.11218527*. [cit. 2025-04-03].

TITAN, 2023. The Text Titan I (Super model). *Transkribus.* Online. 5. 4. 2023. Dostupné z: *https://app.transkribus.org/models/text/51170. [cit. 2025-04-03]*.

ZAVŘELOVÁ, Alžběta, 2020. Projekt PERO – OCR pro historické texty. *Duha: Informace o knihách a knihovnách*. Online. Roč. 34, č. 4. ISSN 1804-4255. Dostupné z: *http://duha.mzk.cz/clanky/projekt-pero-ocr-pro-historicke-texty*. [cit. 2025-04-03].

ŽABIČKA, Petr, 2023. Implementácia umelej inteligencie ako odpoveď na nové výzvy inovatívnych digitálnych služieb. Rozhovor: Petr Žabička – Tomáš Fiala. Online. *ItLib, Informačné technológie a knižnice*. Špeciál 2/2023. Dostupné z: *http://doi.org/10.52036/1335793X.2023.SC2.5-12*. [cit. 2025-04-03].

KATUŠČÁK, Dušan; POHLOVÁ, Klára; NĚMEC, Lukáš; ŘÍHA, Vojtěch. Pokrok v transkripci historických rukopisných dokumentů. *Knihovna: knihovnická revue*. 2025, roč. 36, č. 1, s. 5–30. ISSN 1801-3252.