Martina Dvořáková – Michal Hradiš – Petr Zabička
– Jan Kohút – Martin Kišš – Karel Beneš

VYUŽITÍ PERO OCR PŘI PŘEPISU RUKOPISŮ

ÚVOD

V Česku i ve světě paměťové instituce věnují významné prostředky na digitalizaci písemného kulturního dědictví, bezpečné a dlouhodobě udržitelné uchovávání digitalizátů a v neposlední řadě také na efektivní zpřístupnění digitalizátů, ať už lokálně nebo i vzdálenou formou. Jedním z důležitých aspektů efektivního zpřístupnění je možnost fulltextového vyhledávání v takových sbírkách, které ale vyžaduje jako základní předpoklad kvalitní přepis obsahu. V současnosti není problém kvalitní přepis automaticky vytvořit pro velkou část tištěných dokumentů, ale pro většinu ručně psaných digitalizátů se přepis jejich obsahu v současnosti nevytváří. Kvalitní přepis obsahu je také předpokladem pro další badatelské aktivity, které využívají automatické zpracování obsahu. Jedná se například o kvantitativní i kvalitativní korpusové analýzy (například v *Sketch Engine*), automatickou analýzu témat, analýzu diskurzu nebo synchronní a diachronní jazykovou analýzu. V neposlední řadě je také třeba zmínit využití textového přepisu pro zlepšení čitelnosti, nejen u dokumentů rukopisných, ale i některých tištěných (např. dokumenty tištěné frakturou a švabachem), což tyto dokumenty otevírá dalším zájemcům a obecně zefektivňuje jejich čtení. Textový přepis je možné využít i pro strojový překlad do dalších jazyků, čímž se zpřístupněné dokumenty otevírají celému světu.

Vytvářet dostatečně kvalitní přepisy ručně psaných dokumentů bylo ještě nedávno možné pouze manuálně, ale v posledních letech vznikají automatické nástroje, které umožňují jak čistě automatický přepis, tak umožňují propojit automatické zpracování s manuální kontrolou podle požadované úrovně přesnosti přepisů. V českém prostředí se jedná například o nástroje projektu PERO [2, 9], které zahrnují jak samotné softwarové jádro automatického přepisu, tak i natrénované modely pro písma relevantní v českém prostředí a webovou aplikaci, ve které jsou tyto modely volně dostupné a která poskytuje i efektivní uživatelské rozhraní pro manuální korekce automatických přepisů. Projekt PERO také poskytuje webové API pro hromadné zpracování dokumentů.

Modely poskytované projektem PERO umožňují přepisovat mnoho druhů současného i historického tištěného i ručně psaného písma v mnoha jazycích. Jedná se hlavně o modely pro: 1. tištěné a strojem psané evropské dokumenty včetně velmi nekvalitních předloh, jako jsou například digitalizáty starých noviny skenovaných z mikrofilmů; 2. německou Frakturu; 3. české dokumenty vysázené rodinami fontů švabach, fraktura a podobnými; 4. moderní ručně psané písmo se

zaměřením na český jazyk; 5. německý ručně psaný kurent; 6. český kurent; 7. některá středověká písma.

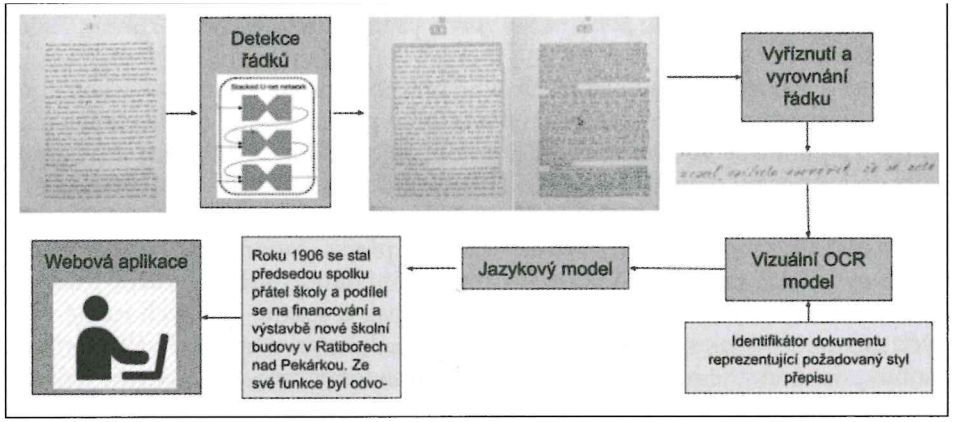
Zmíněné modely dosahují různé úrovně přesnosti přepisu, která závisí na konkrétním použitém modelu, čitelnosti původního písma, kvalitě digitalizátu a také na tom, jak moc se konkrétní dokument odlišuje od dokumentů, které byly použity při trénování daného modelu. Dřívější zkušenosti ukázaly, že například pro nekvalitní a špatně čitelné tištěné dokumenty automatický přepis modely projektu PERO často dosahuje nižší chybovosti než ruční přepis. Automatický přepis je také přímo využitelný pro mnoho účelů při zpracování dobře čitelných ručně psaných českých kronik a podobných dokumentů. Naopak u špatně čitelných ručně psaných dokumentů, jako jsou například deníky vojáků z první světové války a různé osobní poznámky, je někdy ruční kontrola nutná téměř pro všechny typy dalšího využití.

Tento článek shrnuje základní informace o nástrojích projektu PERO a snaží se odpovědět na otázku, jak jsou tyto nástroje využitelné v praxi - tedy jaká je kvalita automatického přepisu textu a v případě využití ručních korekcí, jaká je jejich časová náročnost a jestli takový postup ovlivňuje výslednou kvalitu v porovnání s čistě ručním postupem. Odpovědi na tyto otázky jsou podloženy případovou studií přepisu dokumentů psaných moderním českým písmem a kurentem, která přímo porovnává ruční postup s využitím aplikace PERO-OCR. Článek se dále také opírá o statistiky z webové aplikace PERO-OCR, ze které jsme čerpali jak informace četnosti manuálních oprav, tak informace o jejich časové náročnosti.

PERO OCR

Systém pro automatický přepis písma PERO OCR je implementován jako Python balíček PERO-OCR a jeho zdrojové kódy jsou dostupné pod volnou BSD licencí [1] . Celý systém se skládá z několika kroků, jak je vidět na obrázku 1: 1. detekce řádků a odstavců; 2. vyříznutí řádků; 3. zpracování řádků vizuálním OCR modelem; 4. korekce přepisu pomocí jazykového modelu.

Základem detekce řádků [2] je plně konvoluční neuronová síť, která pro každý pixel vstupního obrázku odhaduje, jestli je daný pixel součástí linky řádku, je koncem řádku nebo je hranicí odstavce. Zároveň tato neuronová síť v každém pixelu odhaduje výšku řádku v pixelech. Sousedící pixely linky řádku jsou pak spojeny a výslednému řádku je přiřazena velikost jako 75% percentil velikostí odhadnutých v pixelech dané linky. Sousední řádky jsou pak spojeny do společného bloku textu, pokud neuronová síť mezi nimi nepredikuje hranici textového bloku. Odstavce jsou seřazeny podle ručně definovaného algoritmu, který se snaží zachovat sloupce textu. Neuronovou síť pro detekci řádků trénujeme na vlastní sadě s ručně zkontrolovanými anotacemi řádků a textových bloků, která zahrnuje široké spektrum dokumentů od novin a moderních tisků přes prvotisky až po ručně psané dokumenty a středověké manuskripty. Tato sada vznikla sjednocením a rozšířením ně-



Obr. 1. Základní řetězec zpracování v PERO OCR.
Zvýrazněné části jsou součástí python balíčku poskytujícího funkcionalitu automatického přepisu.

kolika existujících sad (cBAD [10], IMPACT [11] and BADAM [12]) a je neustále doplňována stránkami, které zpracovávají uživatelé projektu PERO. Část této sady obsahující 683 stránek je zveřejněna jako PERO layout dataset.[1)]

Detekované řádky jsou vyříznuty a vyrovnány do obdélníkových výřezů s konstantní výškou 40 pixelů, které jsou zpracovány vizuálním modelem [9], což jsou neuronové sítě složené z konvolučních vrstev následovaných rekurentními vrstvami. Vizuální modely pro každé 4 pixely vstupního obrázku produkují pravděpodobnost všech možných písmen abecedy (celkem 511 znaků včetně mezery) a prázdného znaku „*blank*". Výsledný textový řetězec může být vytvořen výběrem nejpravděpodobnějšího znaku z každé pozice a vypuštěním znaků „*blank*" (tzv. hladové dekódování, *greedy decoding*). Tyto sítě jsou učeny pomocí chybové funkce CTC [13] na velkých sadách obrázků řádků textu se známými a zkontrolovanými přepisy. Trénování nevyžaduje žádnou znalost pozice jednotlivých písmen v rámci řádku, jen textový řetězec každého řádku. Tabulka 1 shrnuje vizuální modely aktuálně dostupné v rámci aplikace PERO-OCR.

Tab. 1. Přehled modelů dostupných ve webové aplikaci PERO-OCR včetně trénovacích sad, u kterých je vždy v závorce uveden počet řádků. PERO označuje datové sady vytvořené přímo týmem projektu PERO nebo uživateli webové aplikace.

Model	Písmo	Jazyky	Trénovací data
Universal Printed	Antiqua, cyrilice, řecké písmo	Většina evropských	Impact (1.23M) [7] PERO (278k)
Universal Handwritten	Moderní písmo a částečně kurent	Primárně čeština s menší přesností ostatní evropské jazyky	READ (181k) [3] Č. dopisy (87k) [8] Bentham (10k) [4] PERO (390k)
Kurrent	Kurent	Němčina, Čeština	READ (181k) [3] PERO (14k)
German Fraktur	Fraktura	Němčina	DTA (2M)[2)]
Czech fraktur	Fraktura, švabach	Čeština, Němčina	DTA (2M)[3)] PERO (101k)
Medieval Manuscript			Parzival (4.2k) [6] Hab.de (1.3k) 3 Saint Gall (1.3k) [5] PERO (8k)

Různí uživatelé mohou přepisovat stejný dokument různými způsoby. Přepisy se mohou lišit ne jenom díky chybám, ale také díky různorodosti přepisovacích stylů např. dlouhé s může být přepsáno pomocí příslušného UTF-8 symbolu „ſ" nebo standardního znaku „s". V rámci velkých veřejně přístupných systémů je problém různorodosti přepisovacích stylů nevyhnutelný. Běžné neuronové sítě učené pomocí chybové funkce CTC se nejsou schopny naučit přepisovat nekonzistentně přepsané znaky konzistentním způsobem (pokud je přepisovací styl nezávislý na vstupu). Naivním řešení by mohlo být sjednocení přepisovacích stylů uživatelů (zavedení pevně daných přepisovacích stylů). Toto řešení je nežádoucí, jelikož omezuje přepisy na konkrétní styl, tzn. uživatel je nucen přepisovat v rámci povolených stylů a zároveň systém umožňuje přepisovat pouze pomocí těchto stylů. Řešením v rámci projektu PERO je využívání TS-Net [9], což jsou neuronové sítě, které jsou schopné se naučit různé styly přepisu a poté se mezi nimi

[1)] PERO layout dataset: https://www.fit.vut.cz/person/ikodym/pero_layout.zip

[2)] https://www.deutschestextarchiv.de/

[3)] https://diglib.hab.de/

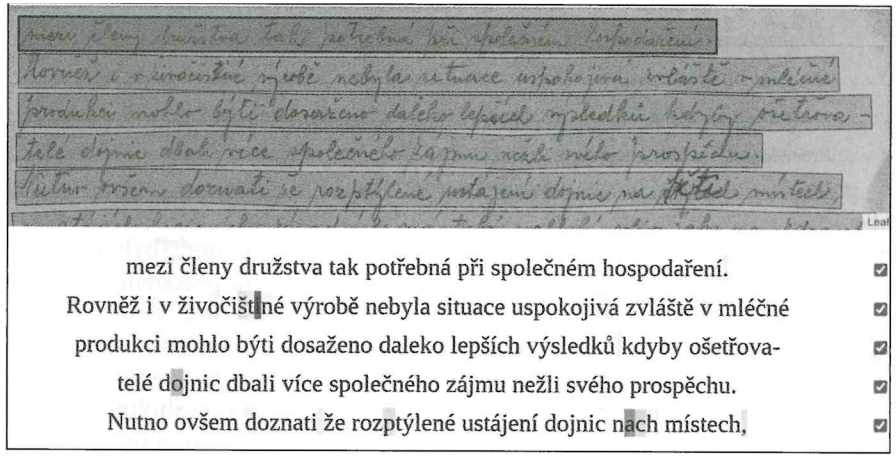
i přepíná. Kromě standardního obrazového vstupu přijímá TS-Net identifikátor přepisovacího stylu (TSI), na základě něhož přepisuje příslušným stylem. TS-Net nevyžaduje žádnou explicitní znalost o vlastnostech přepisovacího stylu, pouze předpokládá, že trénovací data označené pomocí jednoho TSI jsou přepsané konzistentním stylem. Síť se učí přepisovací styly pomocí adaptivní instance normalizace, jejíž parametry jsou podmíněny TSI. TS-Net navíc umožňuje rychlou adaptaci na nový přepisovací styl za pomocí několika řádků. V aplikaci PERO-OCR je většina modelů připravených ve dvou nastaveních: 1. věrné zachování historických znaků a 2. převod do moderní podoby (například i s rozlišením „I" a „J" ve starých německých textech). Každý dokument je také možné přepsat po korekci jeho části. V takovém případě systém automaticky odhadne požadovaný styl přepisu podle již zkontrolovaných řádků daného dokumentu.

Vizuální modely sice již zachycují znalost o jazyce, ale ta není dokonalá a hladové dekódování vybírá jednotlivé znaky přepisu nezávisle na okolních znacích. Pro zvýšení přesnosti přepisu proto používáme tzv. jazykové modely (*language models*), které modelují pravděpodobnost výskytu různých sekvencí písmen v jazyce. Jazykové modely mohou zvýšit přesnost přepisu hlavně u hůře čitelných textů, a pokud se text daného dokumentu příliš neodlišuje od standardu (tyto modely mají například tendenci opravovat překlepy a jiné chyby původního textu). Při použití jazykového modelu využíváme alternativní dekódovací strategii, tzv. prefixové hledání (*prefix search*), kdy pro různé varianty přepisu bereme v potaz všechna jejich možná zarovnání na vstupní obrázek, a jejich vizuální pravděpodobnost kombinujeme právě s pravděpodobností, kterou jim přiřazuje jazykový model. Například v tiscích horší kvality často splývá vizuální podoba písmen o/e/c, ale jazykový model zpravidla dokáže na základě kontextu vybrat správné písmeno s velkou jistotou. V projektu PERO máme nasazeny jazykové modely pro češtinu, němčinu, angličtinu a arabštinu. Tyto modely jsou trénované na velkých korpusech. Zpravidla kombinujeme současné webové korpusy s historickými zdroji a dokumenty uživatelů PERO-OCR.

KOREKCE AUTOMATICKÉHO PŘEPISU S VIZUALIZACÍ NEJISTOTY PÍSMEN

Webová aplikace PERO-OCR[4] poskytuje uživatelské rozhraní pro efektivní manuální korekce automatického přepisu textu (viz Obr. 2). Základem efektivity tohoto rozhraní jsou tři principy: 1. snadná vizuální identifikace potencionálně chybných pasáží přepisu; 2. minimalizace uživatelských akcí ať už myší nebo klávesnicí; 3. udržování prostorové blízkosti přepisu a zdrojového obrazu pro minimalizaci pohybu učí.

[4] Webová aplikace PERO-OCR: https://pero-ocr.fit.vutbr.cz/



Obr. 2. Uživatelské rozhraní pro korekce automatického přepisu v aplikaci PERO-OCR. Nahoře je zobrazen zdrojový obrázek s řádky podbarvenými podle pravděpodobnosti chyby. Dole jsou přepisy s podbarvenými písmeny podle pravděpodobnosti chyby. Po kontrole a potvrzení přepisu jsou zdrojový řádek i textový řádek přepisu podbarveny.

Aplikace zvýrazňuje červenou barvou písmena, kterými si automatický systém není jistý, přičemž intenzita podbarvení odpovídá této nejistotě. Pokud je při automatickém zpracování použit jazykový model, jsou zvýrazněna i písmena, u kterých se liší predikce jazykového a vizuálního modelu. Zároveň jsou řádky ve zdrojovém obrázku podbarveny na škále od modré po červenou podle pravděpodobnosti chyby. Navíc rozhraní poskytuje i tlačítko, které vyhledá následující řádek s pravděpodobnou chybou přepisu – to je užitečné hlavně u velkoformátových tisků s velkým množstvím řádků a s nízkou chybovostí přepisů.

Potřeba uživatelských akcí je minimalizovaná automatickým propojením pohledu na zdrojový obrázek s řádky přepisu. Podle pozice kurzoru v přepisu se automaticky zobrazuje odpovídající řádek zdrojového obrázku a u dlouhých řádků, které se nevejdou na obrazovku, se takový řádek automaticky posunuje podle pozice kurzoru. To umožňuje uživateli pohybovat se pouze v přepisu bez nutnosti posunovat obrázek. Propojení funguje i obráceně – po výběru řádku ve zdrojovém obrázku se automaticky zobrazí odpovídající řádek přepisu.

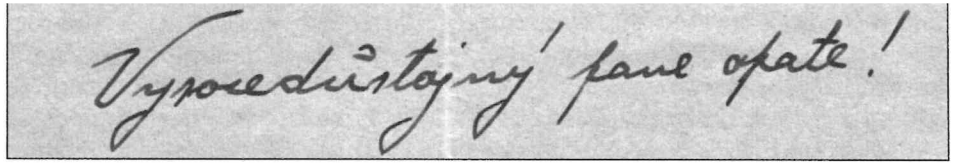
Při kontrolách přepisů musí uživatel často očima přeskakovat mezi zdrojovým obrázkem a přepisem. Rychlost a přesnost těchto sakád je přímo závislá na vzdálenosti, kterou musí překonávat. Zároveň, pokud jsou pozice dostatečně blízké, mohou být současně udržovány v zorném poli, což ještě výrazněji zpřesňuje pohyb očí během sakád. PERO-OCR těchto vlastností lidského vidění využívá a zobrazuje přepis a řádek zdrojového obrázku pod sebou a zároveň co nejblíže se zachováním dostatečného kontextu jak v obrázku, tak v přepisu.

PŘÍPADOVÁ STUDIE VYUŽITÍ PERO OCR PŘI PŘEPISU HISTORICKÝCH DOKUMENTŮ

V rámci praxí studentů Filozofické fakulty Masarykovy univerzity byl realizován experiment, jehož cílem bylo zhodnotit práci při přepisování rukopisných ukázek s aplikací PERO-OCR v porovnání s manuálním přepisem v textovém editoru. Praxe probíhaly během roku 2022 v rozmezí měsíců března a května, samotný experiment poté v rámci měsíce dubna téhož roku. Pro experiment byly vybrány tři studentky rozdílných oborů společně se dvěma odbornými pracovníky z Moravské zemské knihovny a z Fakulty informačních technologií Vysokého učení technického. Každý z účastníků měl za úkol zpracovat určité množství ukázek rukopisného písma, rozděleného podle jejich kompetentnosti, pořízených z digitální knihovny Slovenské národní knihovny [1], která disponuje rozsáhlým množstvím rukopisů různých pisatelů v českém jazyce. Ukázky zpracované praktikantkami byly rozděleny do dvou skupin po jedenácti, z nichž každá obsahovala devět ukázek moderních rukopisů převážně z dvacátého století a čtyři ukázky kurentu ze století devatenáctého. Odborní pracovníci měli sice stejné zadání jako studentky, ovšem jejich úkolem bylo zpracovat odlišné množství ukázek. Odborný pracovník z FIT VUT se v první skupině potýkal celkem s osmi ukázkami moderního rukopisného písma, ve druhé pak s devíti. Ani v jedné skupině však nenarazil na kurent. Největší množství ukázek měl za úkol zpracovat odborný pracovník z paměťové instituce. Každá skupina obsahovala celkem třináct ukázek, z nichž tři byly ukázkami kurentu. Hlavním úkolem účastníků bylo první skupinu ukázek zpracovat v aplikaci PERO-OCR a druhou v libovolném textovém editoru. Během této činnosti si participanti zaznamenávali čas v minutách, který nad jednotlivými ukázkami strávili, do sdíleného souboru v aplikaci Sheets služby Google Disk. Vytvořené přepisy pak byly důkladně znovu zkontrolovány, aby bylo možné vyhodnotit jejich chybovost.

Výsledky této studie jsou shrnuty v tabulce 2. Rychlost přepisu byla výrazně a systematicky vyšší při použití aplikace PERO-OCR (zrychlení přibližně 1,75×). Je zajímavé, že toto zrychlení bylo obdobné u moderního písma i u kurentu, přičemž chybovost automatického přepisu (a tedy i počet manuálních korekcí) byla u kurentu výrazně vyšší (2,7×). Chybovost výsledných přepisů byla u kurentu vyšší při použití aplikace PERO-OCR (1,5×), ale u moderního písma byla naopak nižší (0,77×). Vzhledem k celkově nízké chybovosti (malému počtu individuálních chyb ve vyhodnocovaných stránkách hlavně u kurentu) a faktu, že velká část chyb neplynula z nepozornosti (překlepy), ale z například z nepochopení přesných instrukcí, jak je dále znázorněno ukázkami, není možné z těch čísel vyvozovat, že by některý z uvedených přístupů obecně vedl k nižší chybovosti přepisů.

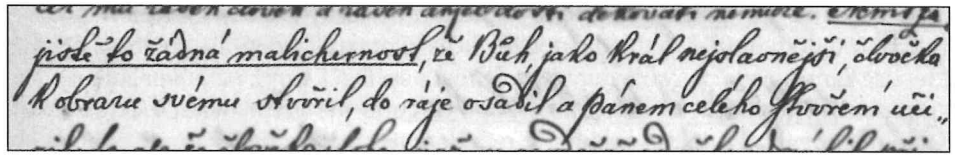
Hlubší analýza chyb ukázala, že participanti se při přepisování ukázek v libovolném nástroji potýkali se stejnými problémy. Při přepisování docházelo často k záměně písmen ve slovech (viz Obr. 3.). U přepisů v textovém editoru několikrát došlo k vynechání celého řádku původního textu (viz Obr. 4.). Častým jevem byla



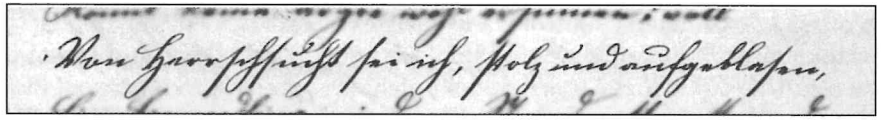
Obr. 3. V oslovení „*Vysocedůstojný pane opate!*“ došlo k záměně písmen „s“ a „p“, kdy písmeno „s“ bylo zaměněno za „p“ a písmeno „p“ za „f“.
Participant přepsal oslovení následovně „*Vypocedůptojný fane ofate!*“.

také záměna interpunkčního znaménka rozdělujícího slova na konci řádku za dnes užívaný spojovník. V neposlední řadě u kurentních textů docházelo k náhradě „kulatého s“ na „dlouhé s“ (viz Obr. 5.).

Pro hlubší pochopení osobních zkušeností byl praktikantkám po dokončení úkolů rozeslán krátký dotazník skládající se ze sedmi uzavřených a jedné otevřené otázky. Odborní pracovníci do dotazníkového šetření nebyli zahrnuti, jelikož patří do vývojového teamu. U uzavřených otázek praktikantky bodovaly na škále od



Obr. 4. Při přepisu textu byl vynechán první řádek.
Na konci druhého řádku je slovo „*učiniſ*“ rozděleno párovým interpunkčním znaménkem.
Participant však v přepisu použil k rozdělení slova dnes používaný spojovník.



Obr. 5. Participant při přepisování původního textu „*Von Herrſchfucht ſei ich, ſtolz und aufgeblaſen*“ nahradil „*dlouhá s*“ kulatými.
Při finální kontrole byl tento fakt zakomponován do výsledné statistiky chybovosti při přepisování textů pouze v případech, kdy si participant tento způsob přepisování zvolil sám, ovšem důsledně je poté nedodržoval.

1 do 5. U první až třetí otázky představovalo číslo jedna nejlepší možné hodnocení a číslo pět nejhorší možné hodnocení. U otázek 4–7 bylo bodování pozměněno a to tak, že čísla jedna až dva značila preferenci přepisování dokumentů v textovém editoru a čísla čtyři až pět v aplikaci PERO-OCR, číslo tři pak neutrální postoj. Výsledky dotazníkového šetření dopadly následovně. V otázce užitečnosti nápovědy při přepisů textů panovala u respondentů přímá shoda. Na škále od 1 do 5 se všechny odpovědi zastavily u čísla 1, což lze interpretovat jako maximální využití nástroje. Z hlediska řízení se odhadem nejistoty textu pomocí červeného podkreslení bylo zjištěno, že dvě respondentky se spíše řídily odhadem nejistot, zatímco zbylá zastává k otázce neutrální postoj. V případě hodnocení uživatelského prostředí se setkáváme v podstatě s identickými výsledky kromě první položky, kde v otázce naučitelnosti dvě respondentky hodnotí na stupnici číslem 2, zatímco jedna zvolila nejlepší možné hodnocení. U efektivnosti, zapamatovatelnosti, designu a satisfakce se již respondentky jednohlasně shodují na nejlepším možném hodnocení. Následuje pět otázek, jež bylo zaměřeno na porovnání aplikace PERO-OCR s libovolným textovým editorem. Hned u čtvrté otázky se odpovědi respondentek rozcházely. Dvě ze tří preferovaly přepisování textů spíše v aplikaci, třetí hodnotila aplikaci jako nejlepší řešení pro textový přepis. Experiment dále ukázal, že se respondentkám nejlépe četly opravované ukázky v aplikaci PERO-OCR. Stejný výsledek byl zaznamenán i u následující otázky mířené na chybovost přepisů. Během oprav textů využily dvě praktikantky český či cizojazyčný slovník, zatímco třetí do pomůcek nahlédla pouze občas. Dotazník uzavírala otevřená otázka, která se zaměřovala na celkovou spokojenost s aplikací, popř. na její problémy. Respondentky hodnotily aplikaci jako nápomocnou při přepisování a četbě textů. Dále oceňovaly její jednoduchost a ovladatelnost. Negativa naopak spatřovaly v tabulce znaků, které vytýkaly úzký výběr speciálních znaků.

PŘÍKLADY ZE SKUTEČNÉHO POUŽITÍ PERO-OCR

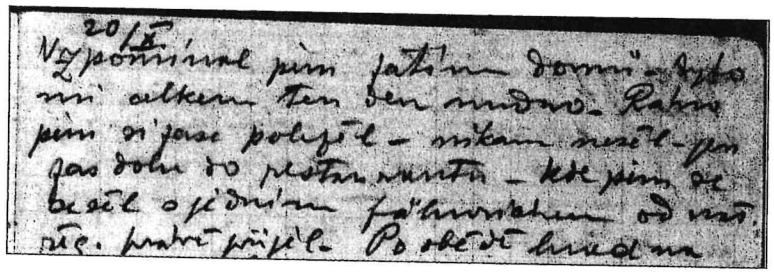
V aplikaci PERO-OCR bylo již zpracováno více než 3 470 dokumentů a z toho bylo u 1 170 dokumentů zkontrolováno více než 10 řádků, u 650 dokumentů přes 100 řádků a u 200 dokumentů více než 1 000 řádků. Tabulka 2 zobrazuje informace o šesti českých ručně psaných dokumentech zpracovaných pomocí PERO-OCR v posledních dvou letech s nejvyšším počtem zkontrolovaných řádků. První tři jsou různé sbírky výňatků z kronik, které jsou většinou dobře čitelné a pro které jsou modely PERO dobře připravené. Následující tři ručně psané deníky vojáků z různého období, kde Vojenský deník 1916 je již příkladem obtížněji čitelného dokumentu (viz Obr. 6). Posledním dokumentem jsou tištěné *Plzeňské Listy* z roku 1898 digitalizované z kvalitního mikrofilmu. Tento dokument byl dříve přepsán za účelem zhodnocení chybovosti modelu pro tištěné dokumenty a je zástupcem dokumentů, pro které daný model funguje velmi kvalitně.

Tab. 2. Chybovosti a rychlosti přepisu ze skutečného použití aplikace PERO-OCR. Tabulka obsahuje šest dokumentů zpracovaných v posledních dvou letech s největším počtem zkontrolovaných řádků. Procento úprav znaků nemusí přímo odpovídat „skutečné" chybě automatického přepisu, ale je to pouze skutečné množství úprav uživatelů aplikace (počítáno pouze z opravených řádků).

Dokument	Počet řádků	Procento úprav znaků	Zkontrolovaných znaků za minutu
Obecní kroniky 1910–1970	29 160	2,3 %	165
Školní a obecní kroniky 1930–1970	16 536	1,6 %	158
Kroniky první pol. 20. stol.	10 286	1,7 %	197
Vojenský deník 1909	7 630	1,2 %	220
Vojenský deník 1968	6 105	3,7 %	119
Vojenský deník 1916	5 513	6,6 %	98
1898 Plzeňské listy (tisk)	555	0,1 %	475

PŘÍNOS APLIKACE

Prvotním impulzem pro vytvoření systému PERO byla nespokojenost se schopnostmi existujících OCR systémů, které jsou vynikající při zpracování současných dokumentů, ale měly značné problémy při zpracování starších tisků a zcela absentovala možnost přepisu rukopisů. Systém PERO je již v Moravské zemské knihovně v Brně integrován do digitalizačního systému ProArc a v budoucnu by měl být zapojen i do digitalizační linky Národní digitální knihovny. Systém PERO umožňuje výrazně zlepšit kvalitu přepisu starých novin, tištěných dokumentů v Digitální knihovně MZK a umožňuje zajistit i textový přepis již digitalizovaných ale dosud do textové podoby nepřevedených dokumentů. Obdobně může systém PERO obohatit i další digitální knihovny, revoluci pak může způsobit při zpřístupnění archiválií, kde se zatím s možností využití automatického textového přepisu právě kvůli charakteru zpřístupněných dokumentů nepočítalo.



Obr. 6. Ukázka z dokumentu Vojenský deník 1916

ZÁVĚR

Nástroje projektu PERO umožňují jednotlivcům i organizacím automaticky přepisovat široké spektrum dokumentů včetně starších tisků a rukopisů. Experiment prezentovaný v tomto článku, který přímo porovnává rychlost a kvalitu manuálního přepisu s poloautomatickým přepisem v aplikaci PERO-OCR, ukazuje na výrazné zrychlení bez negativního vlivu na kvalitu výsledných přepisů. I přes omezený rozsah tohoto experimentu jsou tyto výsledky jasným indikátorem, že automatický přepis s případnou manuální kontrolou má jednoznačně své místo při práci kolekcemi starých dokumentů. Analýza nejdůkladněji zkontrolovaných dokumentů z webové aplikace PERO-OCR ukázala, že pro široké spektrum dobře čitelných českých rukopisů se chybovost automatického přepisu pohybuje od jednoho do čtyř procent.

Do budoucna by bylo vhodné přesněji zmapovat chybovost automatického zpracování na různých druzích dokumentů, odhadnout míru korelace rychlosti manuálních korekcí s chybovostí automatického přepisu a vyhodnotit další možné aspekty vlivu využití automatických nástrojů na kvalitu výsledného přepisu. Je pravděpodobné, že automaticky generovaná předloha ovlivní jiným způsobem nezkušeného čtenáře a jinak odborníka.

Projekt PERO (Pokročilá extrakce a rozpoznávání obsahu tištěných a rukou psaných digitalizátů pro zvýšení jejich přístupnosti a využitelnosti) vznikl ve spolupráci Vysokého učení technického v Brně a Moravské zemské knihovny v Brně a je financován Ministerstvem kultury ČR jako projekt aplikovaného výzkumu a experimentálního vývoje zaměřený na kulturní a národní identitu (NAKI II) DG18P02OVV055.

Seznam literatury

[1] Slovenská národná knižnica | Digitálna knižnica. Slovenská národná knižnica | Digitálna knižnica [online]. Dostupné z: https://onk.snk.sk/

[2] Oldřich Kodym – Michal Hradiš, *Page Layout Analysis System for Unconstrained Historic Documents*. In: J. Lladós – D. Lopresti – S. Uchida (eds), Document Analysis and Recognition – ICDAR 2021. Lecture Notes in Computer Science, vol 12822. Springer, Cham. ICDAR 2021.

[3] Joan Andreu Sanchez – Veronica Romero – Alejandro H. Toselli – Mauricio Villegas – Enrique Vidal, *ICDAR2017 competition on handwritten text recognition on the READ dataset*. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pages 1383–1388. IEEE.

[4] Joan Andreu Sanchez – Veronica Romero – Alejandro H. Toselli – Enrique Vidal, *ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS)*. In: 2014 14th International Conference on Frontiers in Handwriting Recognition, pages 785–790. ISSN: 2167–6445.

[5] Andreas Fischer – Volkmar Frinken – Alicia Fornés – Horst Bunke, *Transcription Alignment of Latin Manuscripts using Hidden Markov Models*. In: Proc. 1st Int. Workshop on Historical Document Imaging and Processing, pages 29–36, 2011.

[6] Andreas Fischer – Andreas Keller – Volkmar Frinken – Horst Bunke, *Lexicon-Free Handwritten Word Spotting Using Character HMMs*. In: Pattern Recognition Letters, Volume 33(7), pages 934–942, 2012.

[7] Christos Papadopoulos – Stefan Pletschacher – Christian Clausner – Apostolos Antonacopoulos, *The IMPACT dataset of historical document images*. In: Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, HIP '13, pages 123–130, New York, NY, USA, August 2013. Association for Computing Machinery.

[8] Zdeňka Hladká, a kol., *111 let českého dopisu v korpusovém zpracování*. Brno 2013.

[9] Jan Kohút – Michal Hradiš, *TS-Net: OCR trained to switch between text transcription styles*. In: International Conference on Document Analysis and Recognition (pp. 478–493). Springer, Cham, 2021.

[10] Markusi Diem – Florian Kleber – Robert Sablatnig – Basilis Gatos, *cBAD: ICDAR2019 competition on baseline detection*. In: ICDAR (2019).

[11] C.Papadopoulos – S. Pletschacher – C. Clausner – A. Antonacopoulos, *The IMPACT dataset of historical document images*. In: Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing - HIP 2013 (2013).

[12] B. Kiessling, – D. S. B. Ezra – M. T. Miller, *BADAM*. In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing - HIP 2019 (2019).

[13] Alex Graves – Santiago Fernández – Faustino Gomez – Jürgen Schmidhuber,