

OPTICKÉ ROZPOZNÁVANIE ZNAKOV V KONTEXTE DIGITALIZÁCIE KOREŠPONDENCIE SLOVENSKÝCH VZDELANCOV V OBDOBÍ 1500 - 1800¹

Veronika Murgašová

Abstrakt

Optické rozpoznávanie znakov (OCR) je dôležitou súčasťou digitalizačného procesu. Výrazne ovplyvňuje prácu s finálnym produktom digitalizácie. Článok bližšie špecifikuje systém rozpoznávania znakov a osobitosti pri digitalizácii rukopisných materiálov, a to konkrétne korešpondencie slovenských intelektuálov v období rokov 1500 – 1800. Problematika súvisí s riešeným projektom APVV-15-0554 Intelektuálne dedičstvo a vedecká komunikácia 1500 - 1800 so slovenskými vzťahmi ako súčasť európskej histórie a identity.

Kľúčové slová: digitalizácia, optické rozpoznávanie znakov, vedecká komunikácia

Abstract

Optical character recognition (OCR) is an important part of digitisation process. It significantly affects work with the final product of digitisation. The article specifies the system of character recognition and specialty in the digitisation of handwritten materials, namely the correspondence of Slovak intellectuals in the period 1500 - 1800. The theme of article related to the solved project Intellectual heritage and scientific communication in 1500 -1800 with Slovak relations as part of European history and identity.

Key words: digitisation, optical character recognition, scientific communication

V rámci projektu APVV-15-0554 s názvom Intelektuálne dedičstvo a vedecká komunikácia 1500-1800 so slovenskými vzťahmi ako súčasť európskej histórie a identity, sa v druhej etape riešiteľský kolektív zaoberá aj digitalizáciou a laboratórnou analýzou nosičov korešpondencie slovenských vzdelancov. Jedným z cieľov projektu sú experimenty zamerané na zvýšenie kvality digitalizácie a rozpoznávania znakov rukopisných dokumentov.

¹ Príspevok je výstupom a súčasťou výskumu podporeného Agentúrou pre vedu a výskum (APVV) v rámci projektu *Intelektuálne dedičstvo a vedecká komunikácia 1500 – 1800 so slovenskými vzťahmi ako súčasť európskej histórie a identity* (APVV-15-0554).

Digitalizácia rukopisov, teda aj predmetnej korešpondencie je stále veľkou výzvou pre všetky pamäťové inštitúcie, a to najmä v oblasti automatického rozpoznávania znakov. Historické materiály, ktoré vstupujú do procesu digitalizácie sú rôznorodé. Listy ako zdroj informácií majú po formálnej stránke svoje špecifiká, na ktoré je potrebné v jednotlivých procesoch digitalizácie prihliadať.

Prvým krokom digitalizácie je skenovanie. Pri korešpondencii z rokov 1500-1800 ide o manuálne skenovanie, ktoré je časovo náročnejšie. Automatické skenovanie je buď nemožné alebo neefektívne. Akákoľvek snaha o automatizáciu môže byť bez predchádzajúcich konzervátorských a reštaurátorských zásahov veľkou hrozbou pre samotný dokument alebo pre skenovacie zariadenie, ktoré sa zanáša čiastočkami uvoľňujúcimi sa z nosiča.

Kvalitný sken dokumentu ovplyvňuje výraznou mierou následný postprocesing a rozpoznávanie znakov. Odporúčané rozlíšenie pre skenovanie rukopisov je minimálne 400 dpi. Skenované súbory sú ukladané vo formáte tiff, ktorý vďaka svojej bezstratovej kompresii zabezpečuje vyššiu kvalitu obrázkov.

Následné kroky v úprave obrazu sa vo veľkej miere zhodujú s akýmkoľvek iným typom dokumentov. V tejto časti procesu je najdôležitejšia snaha odstrániť všetky nedostatky skenov (zakrivenie, presvitanie, kontrast a pod.).

Rozpoznávanie znakov /OCR, ICR, HTR/²

Najvýraznejšie rozdiely pri digitalizácii korešpondencie slovenských vzdelancov vidíme pri rozpoznávaní znakov. Pre tlačene publikácie je na trhu množstvo kvalitných nástrojov, ktoré sa neustále vyvíjajú a ponúkajú kvalitné výstupy. Pri rukopisných dokumentoch však ide o oveľa komplikovanejší proces.

Optické rozpoznávanie znakov (optical character recognition, resp. optical character reader) je veľmi zjednodušene prevod obrazu na text. Je to technológia, ktorá umožňuje konvertovať rôzne typy dokumentov, napr. naskenované papierové dokumenty, súbory PDF alebo snímky vytvorené digitálnym fotoaparátom, do upravovateľných a prehľadateľných údajov. V prípade skenov papierových dokumentov skener nedokáže previesť obraz tak, aby bolo možné ho upravovať napríklad vo Worde. Dokáže len vytvoriť obrazovú kópiu dokumentu,

² OCR – optical character recognition, ICr – Intelligent character recognition, HTR –Handwritten document recognition

ktorá je len súbor čiernych a bielych bodov alebo farebných bodov, ktoré sú zobrazené do mriežky (rastrový, resp. bitmapový obraz). Aby bolo možné vydolovať a použiť údaje z naskenovaných dokumentov, digitálnych fotografií alebo obrázkových súborov PDF, je potrebný práve určitý druh softvéru OCR, ktorý vyextrahuje písmená z obrázka, znak po znaku a prevedie ich na slová, slová na vety, čo následne umožňuje úpravu obsahu originálneho dokumentu.³

Systém práce programu je členený do viacerých etáp. V prvej analyzuje štruktúru snímky dokumentu. Program rozdelí skenovanú stranu do prvkov, ako sú bloky textu, tabuliek, obrázkov a pod. Čiary sú rozdelené do slov a následne do znakov. V druhej etape takýmto spôsobom vyčlenené znaky porovná so súborom vzorových snímok. Výsledkom porovnania je návrh určitého počtu hypotéz o tom, čo znak vlastne predstavuje. V poslednej fáze na základe týchto hypotéz program analyzuje rôzne varianty rozdelenia čiar do slov a slov do znakov. Po spracovaní veľkého počtu predbežných hypotéz program nakoniec rozhodne, ako má vyzeráť rozpoznaný text.⁴

Úlohou operátora digitalizácie je skontrolovať či sa program rozhodol správne. Žiadny program neposkytuje 100 percentné automatické rozpoznávanie znakov, a preto je potrebná následná kontrola a prípadne úpravy, respektíve zásahy v prípade, že program nedokáže niektorý znak rozoznať.

Pri rozpoznávaní znakov sa využívajú tzv. neurónové siete. Ide o siete, ktoré napodobňujú fungovanie ľudského mozgu. Využívajú sa najmä v oblastiach, kde je potrebné, aby sa systém aktívne učil a tým poskytoval relevantné výsledky. Neurónové siete sa využívajú najmä pri rozpoznávaní rukopisov. Systém sa postupne učí rozpoznávať rukou písané znaky a na základe naučených prepojení dokáže určiť o aký znak ide.

Možnosti rozpoznávania znakov pre rukopisné dokumenty

Pracovný postup optického rozpoznávania znakov pre historické zbierky je založený na nástrojoch s otvoreným zdrojovým kódom. Open source nástroje umožňujú lepšie

³ ABBY. *Čo je OCR?* [online]. 2018[cit.20.10.2018].Dostupné na: <https://www.abbyy.com/sk-sk/finereader/what-is-ocr/>

⁴ ABBY. *Čo je OCR?* [online]. 2018[cit.20.10.2018].Dostupné na: <https://www.abbyy.com/sk-sk/finereader/what-is-ocr/>

prispôsobenie požiadaviek, najmä pokiaľ ide o výcvik modelových znakov a modelovanie jazyka na jednotlivé projekty.⁵

Pri rukopisných dokumentoch narážajú programy na viaceré problémy, ktoré je možné rozdeliť do troch väčších skupín.

- Problémy s obrazom - Jedným z hlavných problémov je kvalita pôvodného rukopisu a kvality skenovania. Zahŕňa to problémy, ktoré nastávajú pri manipulácii s originálnym dokumentom, ako sú skrútené stránky, rozmazané písmo alebo dodatočne manuálne upravované stránky (napríklad pečiatky alebo ďalšie ručne vpísané poznámky mimo líniu textu).
- Typ písma a rozloženie - Ľudské rukopisy nie je možné spracovať štandardným softvérom OCR. Písma alebo znaky sa líšia pri zmene osoby, okrem písmen sú často nekonzistentné aj medzery medzi. Historické dokumenty často používajú rozdielne a nekonzistentné štruktúry usporiadania textu. Rukopis jednej osoby sa tiež vekom mení, a preto aj učenie programov býva náročné.
- Chýbajúca vedomostná báza - tradičný softvér OCR používa znalostné bázy založené na súčasných slovníkoch a gramatické štruktúry na zlepšenie postupu OCR. Takéto softvéry neposkytujú bázy pre rukopisné dokumenty. Historické rukopisy často nesledujú konkrétne pravopisné štruktúry a pravidlá, teda môžu byť napísané slová v tom istom texte inak.⁶

Nástroj Transcribus

Transkribus je nástroj, ktorý je spravovaný pracovnou skupinou *Digitalizácia a digitálna konzervácia* (DEA) na univerzite v Innsbrucku. Je financovaný Európskou komisiou ako súčasť projektu READ H2020 riešeného v rokoch 2016 – 2019.

Transkribus je komplexná platforma pre automatické rozpoznávanie, prepisovanie a vyhľadávanie v historických dokumentoch.

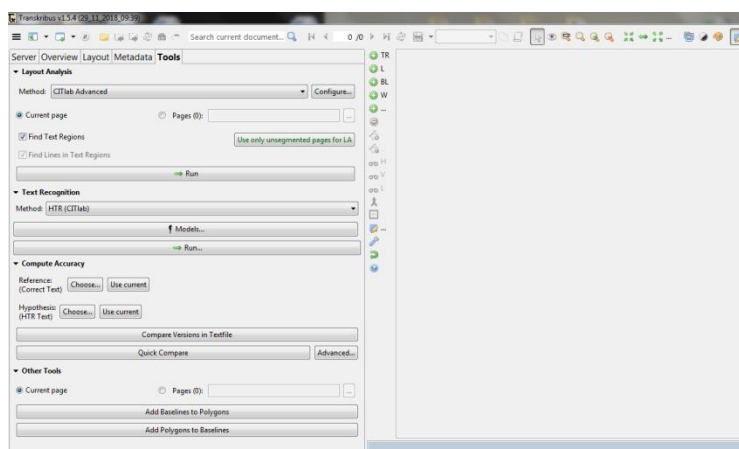
Skladá sa z:

- expertného nástroja (Transkribus – Obr.1)
- webinterface (<http://transkribus.eu/>)
- niekoľko cloudových služieb.

⁵ BLANKE, Tobias a BRYANT, Michael a Mark HEDGES. Ocropodium: Open source OCR for small-scale historical archives. In: *Journal of Information Science*. roč. 38, č. 1, 2012, s.76-86. ISSN 0165-5515.

⁶ SUKANYA, Roy, TRIDIB, Dawn a Deb SUBRATA. A study on handwriting analysis by OCR. In: *International Journal of Scientific and Research Publications*. Vol. 8, Is. 1, 2018, s. 449-450. ISSN 2250-3153

Hlavným cieľom spoločnosti Transkribus je podporovať používateľov, ktorí sa zapájajú do prepisu vytlačených alebo ručne písaných dokumentov, konkrétne vedeckých a výskumných pracovníkov, archívy, ale aj dobrovoľníkov z verejnosti.⁷



Obr. 1 Pracovné prostredie programu Transkribus

Transkribus je stále vo vývoji a väčšina častí softvéru je open source. Vzhľadom na to, že ide o open source ponúka pre inštitúcie veľkú výhodu v podobe minimálnej finančnej záťaže.

Takéto typy nástrojov ponúkajú viacero výhod. Okrem toho, že sú voľne dostupné na používanie, sú niekedy flexibilnejšie a majú nižšie administratívne náklady v porovnaní s porovnateľnými komerčnými produktmi. Pokiaľ ide o schopnosti čistého rozpoznávania textu, výkonnostná medzera medzi komerčnými nástrojmi a najlepšimi dostupnými zdrojmi open source stále existuje, je však stále menšia a menšia.

Nástroj Transkribus je v súčasnosti testovaný v prostredí digitalizačného centra Žilinskej univerzity v Žiline a výstupy s experimentov budú publikované v samostatnej štúdii.

Záver

Cieľom akéhokoľvek rozpoznávania textu v naskenovanom obraze je extrahovať z neho text, ktorý používateľ môže následne ďalej editovať. Rozpoznávanie znakov je preto jedným zo základných predpokladov pre prácu s digitalizovanými materiálmi.

Flexibilita našich digitalizačných nástrojov musí byť priamo úmerná variabilite zdrojových dokumentov. Pracovné postupy, ktoré lepšie zodpovedajú rozličným potrebám vstupných materiálov, umožnia zamestnancom lepšie rozmiestniť svoje zručnosti a vedomosti pri prispievaní k nášmu digitálnemu kultúrnemu a vedeckému dedičstvu.

⁷ DIGITISATION AND DIGITAL PRESERVATION GROUP. *Transkribus* [online]. TRANSKRIBUS Team at University of Innsbruck, 2018[cit.23.11.2018]. Dostupné na: <https://transkribus.eu/Transkribus/>

Zoznam použitej literatúry

ABBY. *Čo je OCR?* [online]. 2018[cit.20.10.2018]. Dostupné na: <https://www.abbyy.com/sk-sk/finereader/what-is-ocr/>

BLANKE, Tobias a BRYANT, Michael a Mark HEDGES. Ocropodium: Open source OCR for small-scale historical archives. In: *Journal of Information Science*. roč. 38, č. 1, 2012, s.76-86. ISSN 0165-5515

SUKANYA, Roy, TRIDIB, Dawn a Deb SUBRATA. A study on handwriting analysis by OCR. In: *International Journal of Scientific and Research Publications*. roč. 8, č. 1, 2018, s. 449-450. ISSN 2250-3153

DIGITISATION AND DIGITAL PRESERVATION GROUP. *Transkribus* [online].

TRANSKRIBUS Team at University of Innsbruck, 2018[cit.23.11.2018]. Dostupné na: <https://transkribus.eu/Transkribus/>