



Záverečná karta projektu

Názov projektu

Evidenčné číslo projektu

APVV-19-0456

Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripce historických rukopisov

Zodpovedný riešiteľ **doc. Mgr. Imrich Nagy, PhD.**

Príjemca **Univerzita Mateja Bela v Banskej Bystrici**

Názov pracoviska, na ktorom bol projekt riešený

Univerzita Mateja Bela v Banskej Bystrici v spolupráci so Štátnou vedeckou knižnicou v Banskej Bystrici

Názov a štát zahraničného pracoviska, ktoré spolupracovalo pri riešení

Na riešení projektu priamo nespolupracovalo zahraničné pracovisko. Univerzita Mateja Bela však v rámci riešenia projektu vstúpila do združenia READ-COOP so sídlom v Innsbrucku (Rakúsko), ktoré je prevádzkovateľom nástroja umelej inteligencie Transkribus a jeho aplikácií.

Udelené patenty/podané patentové prihlášky, vynálezy alebo úžitkové vzory, ktoré sú výsledkami projektu

Hlavnými výsledkami projektu sú dva agregované modely pre automatickú transkripciu slovacikálnych historických dokumentov vytvorené na základe špecifických modelov vytrénovaných počas riešenia projektu pre konkrétne dokumenty:

ŠTANDARDNÁ CITÁCIA DATASETU Slovak Supermodel M1 (SSM1).

KATUŠČÁK, DUŠAN, IMRICH NAGY, PAVOL MALINIAK, ALICA KURHAJCOVÁ, OTO TOMEČEK, PATRIK KUNEC, A MÁRIA BÔBOVÁ. (2024). Slovak Supermodel M1 (SSM1). Dataset. Model ID: 63569. First version, 20240503. Embarged. CC BY-NC-SA. – Reposit: Zenodo. <https://doi.org/10.5281/zenodo.11109087>. Transkribus app. URL:

<https://app.transkribus.org/training/text/63569>. – Notes: Slovak, Latin, Hungarian, Czech languages. Size: 1359 pages Ground Truth (GT); 56713 lines a 333777 words. Train Set: 4,90 % a Validation Set: 5,30 %. – Funding: SKRIPTOR project – Innovative access to the written heritage of Slovakia through a system of automatic transcription of historical manuscripts. APVV-19-0456 SKRIPTOR (2021-2024).

ŠTANDARDNÁ CITÁCIA DATASETU Slovak Supermodel P&T1 (SSPT1).

KATUŠČÁK, DUŠAN, LUCIA NIŽNÍKOVÁ, MICHAELA MIKUŠKOVÁ, NIKOLA HALFAROVÁ, TEREZIE GAJDOŠOVÁ, LENKA MÁLKOVÁ, NIKOL TAUFROVÁ, et al. (2024). Slovak Supermodel P&T1 (SSPT1). Dataset. Model ID: 78289. First version, 20240520. – CC BY-NC-SA. – Zenodo. <https://doi.org/10.5281/zenodo.11218527>.

Transkribus app. URL: <https://app.transkribus.org/training/text/78289>. – Notes: SLO, CZE, UPP, GER languages. - Training Set Size 200 697. CER (Accuracy) 1.00%. – Funding: SKRIPTOR project – Innovative access to the written heritage of Slovakia through a system of automatic transcription of historical manuscripts. APVV-19-0456 SKRIPTOR (2021-2024).

Najvýznamnejšie publikácie (knihy, články, prednášky, správy a pod.) zhrňujúce výsledky projektu – uveďte aj publikácie prijaté do tlače

BÔBOVÁ, Mária. Možnosti využitia platformy Transkribus v prostredí dejín knižnej kultúry. In: Rostislav KRÚŠINSKÝ, ed.: *Bibliotheca antiqua 2023: sborník z 31. konferencie*. Olomouc: Vědecká knihovna v Olomouci, 2023, s. 83-89. ISBN 978-80-7053-343-7.

BÔBOVÁ, Mária. Prepis rukopisných katalógov z územia Slovenska prostredníctvom platformy Transkribus. In: *Knižnica*, 25, 2024, 4.

- v tlači.

KATUŠČÁK, Dušan. Umelá inteligencia pomáha sprístupňovať písomné dedičstvo. In: *Knihovna : knihovnícká revue*. 2022, 33(2), s. 50 – 77, [1,74 AH]. ISSN 1801-3252.

Dostupné na: <https://knihovnarevue.nkp.cz/archiv/dokumenty/2022-2/katuscak.pdf>

KATUŠČÁK, Dušan. Robot s umelou inteligenciou sprístupňuje písomné dedičstvo. *Knižnice 2022 – Priestor pre všetkých* Príspevky z celoslovenskej konferencie konanej 4. – 5. októbra 2022 v Liptovskom Jáne. In: Ondrej Látka, rec. Mgr. Beáta Bellérová, PhD., Mgr. Jana Ilavská, PhD. *Knižnice 2022 – Priestor pre všetkých : Príspevky z celoslovenskej konferencie konanej 4. – 5. októbra 2022 v Liptovskom Jáne*. 1. vyd. Bratislava: Slovenská asociácia knižníc, Bratislava, 2023. s. 22-33. ISBN 978-80-972135-4-1. doi:10.33542/CKK22-4-1. Dostupné na: <https://repo.umb.sk/handle/123456789/413>

KATUŠČÁK, Dušan – NAGY, Imrich (eds.). *Automatická transkripcia slovacikálnych historických dokumentov* [elektronická editovaná kniha]; rec. Milan Konvit, Jan Odstrčilík. - 1. vyd. - Banská Bystrica : Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici, 2022. - [14,2 AH] - ISBN 978-80-557-2020-3. - DOI

<https://doi.org/10.24040/2022.9788055720203>

KATUŠČÁK, Dušan, NAGY, Imrich (eds.). *Automatická transkripcia historických dokumentov : metodická príručka na prácu s platformou Transkribus*. Banská Bystrica : Vydavateľstvo Univerzity Mateja Bela - Belianum, 2023, 156 s. ISBN 978-80-557-2070-8. Dostupné na: <https://repo.umb.sk/xmlui/handle/123456789/412>

[BÔBOVÁ, Mária (7%) - KATUŠČÁK, Dušan (32%) - KUNEC, Patrik (9%) - KURHAJCOVÁ, Alica (5%) - MALINIAK, Pavol (7%) - MIKUŠKOVÁ, Michaela (22%) - NAGY, Imrich (10%) - NIŽNÍKOVÁ, Lucia (6%) - TOMEČEK, Oto (2%)]

KATUŠČÁK, Dušan, NAGY, Imrich (eds.): *Automatická transkripcia historických dokumentov v prostredí webovej aplikácie Transkribus : metodická príručka pre účastníkov workshopu*. verzia Transkribus app. 3.1.0.103. - 1. vyd. - Banská Bystrica : Vydavateľstvo Univerzity Mateja Bela - Belianum, 2024. - 134 s. [5,15 AH]. - ISBN 978-80-557-2143-9. Dostupné na: <https://repo.umb.sk/xmlui/handle/123456789/506>

[BÔBOVÁ, Mária (6,00%) - KATUŠČÁK, Dušan (33,81%) - KURHAJCOVÁ, Alica (5,26%) - MALINIAK, Pavol (8,36%) - MIKUŠKOVÁ, Michaela (24,30%) - NAGY, Imrich (14,47%) - NIŽNÍKOVÁ, Lucia (3,90%) - TOMEČEK, Oto (3,90%)]

MALINIAK, Pavol. Tematizácia osmanského ohrozenia v kazateľskej tvorbe Izáka Abrahamidesa Hrochotského. In: *Kultúrne dejiny*. roč. 13, č. 2, 2022, s. 179 – 198, [1,99 AH]. Dostupné na: <https://doi.org/10.54937/kd.2022.13.2.179-198>

MALINIAK, Pavol. Podoby viacjazyčnosti v rukopisných kázňach Izáka Abrahamidesa Hrochotského = Forms of multilingualism in the manuscript sermons of Isaac Abrahamides Hrochotius. In: *Slavica Slovaca*. - Bratislava : Slovenská akadémia vied, 2024. - ISSN 0037-6787. - Roč. 59, č. 1 (2024), s. 150-158. <https://doi.org/10.31577/SlavSlov.2024.1.15>

MIKUŠKOVÁ, Michaela a NIŽNÍKOVÁ, Lucia. Komenský a Bel v rukách umelej inteligencie. Online. *ITlib: informačné technológie a knižnice*. 2024, č. 1, s. 39-51. ISSN 1336-0779. [cit. 2024-07-23]. Dostupné na: <https://doi.org/10.52036/1335793X.2024.1.39-51>

NAGY, Imrich. Možnosti aplikácie metódy digitálnej transkripcie historických rukopisných textov pri sprístupňovaní archívnych fondov = The Possibilities of application the method of digital transcription of historical manuscript texts in the process of accessing the archival fonds. In: *Slovenská archivistika*. - Bratislava : Ministerstvo vnútra Slovenskej republiky, 2021. - ISSN 0231-6722. - Roč. 51, č. 2 (2021), s. 53-67. Dostupné na:

https://www.minv.sk/swift_data/source/verejna_sprava/odbor_archivov_a_registratur/archivnictvo/slovenska_archivistika/SA%202-2021,%20roc.%2051.pdf

TOMEČEK, Oto. Metales Banskej Bystrice z roku 1820. Reambulácia juhozápadného úseku mestských hraníc spoločných so susedným teritóriom rodiny Radvanských = Metales of the town Banská Bystrica from 1820. Perambulation of the southwest part of town borderline common with neighbouring domain of Radvanský family. In: *Acta historica Neosoliensia* :

vedecký časopis pre historické vedy. - Banská Bystrica : Vydavateľstvo Univerzity Mateja Bela - Belianum, 2021. - ISSN 1336-9148. - Roč. 24, č. 2 (2021), s. 112-133. Dostupné na: <https://www.ahn.umb.sk/tomus-24-num-2-tomecek-o-metales-banskej-bystrice-z-roku-1820-reambulacia-juhozapadneho-useku-mestskych-hranic-spolocnych-so-susednym-teritoriom-rodiny-radvanskych/>

TOMEČEK, Oto. Historická krajina v metáciach – výskumné možnosti a limity. In: Cestou necestou (krajina – priestor – kultúra – osobnosti). Prešov : ŠVK v Prešove, 2022, s. 9 – 25, [0,78 AH]. Dostupné na: <https://repo.umb.sk/xmlui/handle/123456789/367>

TOMEČEK, Oto – NAGY, Imrich. Reambulačný protokol mesta Banská Bystrica z roku 1820 : Edícia prameňa pomocou nástroja na automatickú transkripciu historických dokumentov. Banská Bystrica : Belianum, 2024. [16,4 AH]

- v tlači

Uplatnenie výsledkov projektu

Projekt Skriptor overil funkčnosť a úspešnosť nástroja AI na automatický prepis historických rukopisných a tlačných dokumentov Transkribus pre slovacikálne dokumenty. Výsledkom je 6 špecifických modelov pre konkrétne historické rukopisné a tlačné dokumenty.

Vlastníkom týchto dokumentov sú výskumné a pamäťové inštitúcie: Literárny archív – Slovenská národná knižnica v Martine, Odbor archívov a registratúr Ministerstva vnútra SR v zastúpení Štátneho archívu v Banskej Bystrici a Slovenský banský archív v Banskej Štiavnici ako špecializované pracovisko Slovenského národného archívu v Bratislave, resp. Diecézny archív rímskokatolíckej cirkvi – Biskupstvo Banská Bystrica a ďalšie vedecké knižnice. Uvedení vlastníci dokumentov sú odberateľmi prepisov dokumentov, ktoré poskytli pre potreby projektu vrátane kompletných datasetov vyvinutých modelov automatickej transkripcie. V prípade Literárneho archívu – Slovenskej národnej knižnice toto plnenie vyplývalo aj zo zmluvne ukotvenej spolupráce (Zmluva o spolupráci pri implementácii projektových aktivít uzatvorenej medzi Univerzitou Mateja Bela v Banskej Bystrici ako hlavným riešiteľom projektu a Slovenskou národnou knižnicou v Martine).

Na základe najkvalitnejších prepisov, ktoré vznikli pri vývoji uvedených špecifických modelov, boli ďalej vyvinuté dva nové „supermodely“ transkripcie ako významné a inovatívne výstupy projektu, ktoré umožnia sprístupnenie určitého typu historického rukopisného, resp. strojopisného a tlačného písomného dedičstva. Tieto umožňujú spolupracujúcim pamäťovým inštitúciám a ich vedecko-výskumným pracovníkom získať automatický prepis ďalších historických dokumentov, ktoré vlastnia vo svojich fondoch, prípadne podľa potreby a svojich nárokov, na základe týchto supermodelov vyvinúť vlastné špecifické modely pre konkrétne dokumenty s výslednou nižšou mierou chybovosti. Vďaka tejto kooperácii, ktorá prinesie nové datasety korigované do verzie Ground Truth, bude možné ďalej zdokonaľovať zmienené supermodely, t. j. znižovať ich chybovosť a zvyšovať účinnosť, čím sa zrýchli a zefektívni reálna (obsahová) digitalizácia historických dokumentov z písomného dedičstva SR.

Nevyhnutným predpokladom pre dosiahnutie tohto impaktu výsledkov projektu je osvojenie metodiky a techniky práce pri vývoji modelov, príprave dokumentov na automatickú transkripciu a pri samotnej automatickej transkripcii v prostredí platformy Transkribus. V rámci riešenia projektu sa preto vypracovala metodika na prácu v prostredí klientskeho softvéru aj webovej aplikácie Transkribus, ktorá je k dispozícii pre záujemcov z pamäťových inštitúcií, odborníkov z praxe i zástupcov laickej verejnosti formou open access metodických príručiek. Pre osvojenie postupov práce v platforme Transkribus bude Univerzita Mateja Bela v rámci udržateľnosti projektu organizovať pravidelné workshopy, kde si záujemcovia z odbornej i laickej verejnosti budú môcť prakticky vyskúšať a preveriť možnosti transkripcie historických dokumentov prostredníctvom AI.

Vyvolaným dopadom takto uplatnených výsledkov projektu bude sprístupnenie celých rozsiahlych korpusov historických dokumentov pre ich ďalšiu analýzu v rámci základného historického výskumu, resp. sprístupnenie pre vyhľadávanie a získavanie ďalších informácií záujemcami z radov odbornej i laickej verejnosti.

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v slovenskom jazyku (max. 20 riadkov)

V rámci projektu Skriptor sme overili 6 špecifických modelov a 2 univerzálne supermodely pre automatickú transkripciu historických rukopisných dokumentov slovacikálnej

proveniencie. Overili sme funkčnosť a úspešnosť technológie AI Transkribus vyvíjanej a prevádzkovej konzorciom READ-COOP so sídlom v Innsbrucku (Rakúsko), ktorého členom je aj Univerzita Mateja Bela v Banskej Bystrici. Pri špecifických modeloch dosiahla chybovosť automatického prepisu úroveň od 1 po 7 % na úrovni znakov v závislosti od typu dokumentu. Vďaka týmto modelom boli automaticky prepísané a sprístupnené pre ďalší historický výskum rozsiahle korpusy prameňov (3539 digitalizátov). Funkčnosť univerzálného supermodelu pre historické rukopisy bola overená na náhodnom dokumente s výsledkom chybovosti približne 10 % na úrovni znakov, čo predstavuje výsledok použiteľný pre porozumenie a ďalšiu prácu s textom. Zároveň sme vypracovali metodiku práce nástrojom AI na automatický prepis historických dokumentov Transkribus v prostredí klientskeho softvéru (Transkribus Expert Client) aj webovej aplikácie (App Transkribus). Na základe tejto metodiky sme vytvorili vzdelávací kurz – workshop na otestovanie a osvojenie možností automatickej transkripcie historických dokumentov, ktorý Univerzita Mateja Bela ponúka a bude aj naďalej ponúkať pamäťovým inštitúciám, vedecko-výskumným pracovníkom i záujemcom z praxe (Workshop Automatická transkripcia historických rukopisných a tlačných dokumentov). Hlavnou devízou výsledkov projektu je masívne sprístupnenie obsahu archívnych historických dokumentov pre ďalší historický výskum. Napriek časovej náročnosti vývoja modelov pre automatickú transkripciu sa medzi výsledky projektu môžu zaradiť aj jednotlivé výstupy riešiteľov prezentujúce aj tento aspekt ich výskumu.

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v anglickom jazyku (max. 20 riadkov)

As part of the Skriptor project, we verified 6 specific models and 2 universal supermodels for the automatic transcription of historical manuscript documents of Slovak provenance. We validated the functionality and success of the AI technology Transkribus, developed and operated by the READ-COOP consortium based in Innsbruck, Austria, of which Matej Bel University in Banská Bystrica is a member. For specific models, the transcription error rate ranged from 1 to 7% at the character level, depending on the type of document. Thanks to these models, extensive corpora of sources (3539 digitized items) have been automatically transcribed and made available for further historical research. The functionality of the universal supermodel for historical manuscripts was verified on a random document with an error rate of approximately 10% at the character level, which is a usable result for understanding and further work with the text. We also developed a methodology for working with the AI tool for the automatic transcription of historical documents Transkribus in both the client software environment (Transkribus Expert Client) and the web application (App Transkribus). Based on this methodology, we created an educational course - a workshop to test and master the possibilities of automatic transcription of historical documents, which Matej Bel University offers and will continue to offer to memory institutions, scientific researchers, and interested practitioners (Workshop on Automatic Transcription of Historical Manuscript and Printed Documents). The main asset of the project's results is the massive accessibility of the content of archival historical documents for further historical research. Despite the time-consuming nature of developing models for automatic transcription, the project's outcomes also include individual outputs from the researchers that present this aspect of their research.