Historia Aperta

vol. 54, 2025, pp. 9–33
DOI: https://doi.org/10.26806/hisape.n54.1

Model HTR Latin Sermons 11th–12th C.: Metodologie, transkripce, testování a evaluace

MICHAELA FALÁTKOVÁ
BARBORA KULHOVÁ

Abstrakt | Abstract

Studie představuje metodologii tvorby modelu pro automatické rozpoznávání rukopisných textů (HTR) zaměřeného na latinské homilie z 11.–12. století. Model Latin Sermons 11th–12th C. byl trénován na datech pocházejících z osmi rukopisů uložených v knihovnách pěti evropských zemí, psaných pozdní karolinskou a raně gotickou minuskulou. Text se zaměřuje na přípravu tréninkových dat, volbu transkripčních pravidel a testování vlivu parametrů, jako je rozsah dat, počet tréninkových cyklů a využití base modelu. Výsledný model dosáhl míry chybovosti CER 5,7 % a prokázal vysokou adaptabilitu, čímž otevírá nové možnosti pro využití v rámci digital humanities.

Model HTR Latin Sermons 11th–12th C.: Methodology, Transcription, Testing, and Evaluation

This study introduces the methodology behind creating a Handwritten Text Recognition (HTR) model for Latin homilies from the 11th–12th centuries. The Latin Sermons 11th–12th C. model was trained on data originating from eight manuscripts held in libraries across five European countries, written in late Carolingian and early Gothic minuscule. The article focuses on preparing training data, selecting transcription rules, and testing parameters such as data volume, number of epochs, and the use of a base model. The final model achieved a Character Error Rate (CER) of 5.7 % and demonstrated high adaptability, thereby opening new possibilities for use within Digital Humanities.

Klíčová slova | Keywords

HTR; latinské homilie; středověké rukopisy; Transkribus; Digital Humanities
HTR; Latin homilies; medieval manuscripts; Transkribus; Digital Humanities

Automatické rozpoznávání rukopisných textů (*Handwritten Text Recognition*, dále jen HTR) představuje v oblasti digitálních humanitních věd (*digital humanities*) zásadní nástroj, který významně zefektivňuje práci s historickými prameny. V současné odborné praxi se termín *Optical Character Recognition* (dále OCR) obvykle užívá ve vztahu k tištěným textům, zatímco rozpoznávání ručně psaného písma je označováno jako HTR. V poslední době však dochází ke stírání rozdílů mezi oběma technologiemi zejména v důsledku skutečnosti, že obě dnes často využívají principy hlubokého učení (*deep learning*). V této souvislosti se stále častěji setkáváme s obecnějším pojmem *Automatic Text Recognition* (dále ATR), který lze použít v kontextu rozpoznávání jak tištěného, tak rukopisného materiálu.[1] Zatímco technologie OCR se začala intenzivně rozvíjet již v 70. letech 20. století a dnes je běžnou součástí kancelářských softwarových nástrojů (např. *ABBYY FineReader* či *Adobe Acrobat*), HTR zaznamenalo výrazný pokrok teprve s nástupem moderních metod strojového učení (*machine learning*). Klíčovou roli zde sehrály konvoluční neuronové sítě (*Convolutional Neural Networks*, dále CNN) zaměřené na analýzu vizuálních znaků písma a rekurentní neuronové sítě (*Recurrent Neural Networks*, dále RNN) určené k práci s posloupnostmi znaků. Dalším významným faktorem je využití rozsáhlých jazykových modelů (*Large Language Models*, dále LLM), které výrazně přispívají ke snížení chybovosti automatického přepisu.[2]

Tato studie se zaměřuje na platformu *Transkribus*, především na její webovou aplikaci, která uživatelům zpřístupňuje nástroje pro práci s historickými dokumenty. Projekt *Transkribus* byl spuštěn v roce 2015 evropskou iniciativou *tranScriptorium* a následně dále rozvíjen v projektu *READ* (*Recognition and Enrichment of Archival Documents*), jehož cílem bylo podpořit inovace v oblasti archivnictví. V současnosti je *Transkribus* dostupný jak ve formě starší desktopové aplikace, tak prostřednictvím online rozhraní. Mezi hlavní skupiny uživatelů platformy patří badatelé v oblasti humanitních věd,

1 Melissa TERRAS – Bettina ANZINGER – Andy STAUDER a kol., *The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition*, 2025, https://open-research-europe.ec.europa.eu/articles/5-16; všechny online zdroje byly ověřeny k datu 22. 9. 2025.

2 K problematice technologie OCR a HTR srov. Tobias HODEL, *Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft Anwendung, Einordnung und Methodenkritik*, Historische Zeitschrift 316, 2023, s. 151–180. K úvodu do problematiky HTR srov. Ariane PINCHE – Peter STOKES, *Historical Documents and Automatic Recognition: Introduction*, *Journal of Data Mining & Digital Humanities,* 2024, https://doi.org/10.46298/jdmdh.13247. K digitalizaci srov. Matthew CHRISTY – Anshul GUPTA – Elizabeth GRUMBACH – Laura MANDALL – Richard FURUTA – Ricardo GUTIERREZ-OSUNA, *Mass Digitization of Early Modern Texts With Optical Character Recognition*, *Journal on Computing and Cultural Heritage 11*, 2017, č. 1, s. 1–25, https://doi.org/10.1145/3075645.

archivní instituce a knihovny, odborníci na výpočetní technologie a širší veřejnost.[3]

Výchozím bodem pracovního postupu (*workflow*) v aplikaci *Transkribus* je digitalizovaný dokument. Pro vytvoření automatického přepisu je nezbytné na dokument aplikovat vhodný model pro automatické rozpoznávání rukopisného textu (*Handwritten Text Recognition*, dále HTR). Proces přepisu začíná segmentací, tedy automatickým rozpoznáním rozvržení strany (*automatic layout recognition*), při níž je dokument rozdělen na textové bloky (*text regions*) a dále na jednotlivé řádky (*baselines*). Segmentace může být provedena zároveň s aplikací modelu nebo samostatně, což je užitečné zejména při tvorbě tréninkových dat (*training data*), kdy je nutné dokumenty manuálně transkribovat.[4] Automaticky vygenerovaný přepis lze následně upravovat, doplňovat o barevné značky (*tags*), exportovat nebo využít pro trénink vlastního modelu.[5] Takto vytvořený textový dokument spolu s příslušnou digitální kopií je možné sdílet prostřednictvím platformy *Transkribus Sites*.[6]

Vedle platformy *Transkribus* existují i alternativní softwarová řešení využívající technologii HTR. Jedním z příkladů je *eScriptorium*, projekt vyvíjený od roku 2018 na *École Pratique des Hautes Études* při *Université Paris Sciences et Lettres* (EPHE – PSL). V českém kontextu stojí za zmínku iniciativa *PERO*, která vznikla ve spolupráci Vysokého učení technického v Brně a Moravské zemské knihovny. Tento projekt se zaměřuje na zlepšení přístupnosti digitalizovaných historických dokumentů a umožnění fulltextového vyhledávání v digitálních databázích. Systém *PERO* nabízí možnost automatického přepisu jak odborným institucím, tak i širší veřejnosti prostřednictvím

3 Webová aplikace Transkribus, https://www.transkribus.org/. K problematice Transkribu srov. Guenter MUEHLBERGER a kol., *Transforming scholarship in the archives through handwritten text recognition Transkribus as a case study, Journal of Documentation 75,* 2019, č. 5, s. 954–976, https://www.emerald.com/jd/article/75/5/954/206914/Transforming-scholarship-in-the-archives-through. Joe NOCKELS – Paul GOODING – Sarah AMES a kol., *Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research*, *Archival Science 22*, 2022, s. 367–392, https://doi.org/10.1007/s10502-022-09397-0. K Projektu READ srov. Tobias HODEL, *Lesende Algorithmen: Projekt READ*, *Das Mittelalter*, 2019, https://doi.org/10.1515/mial-2019-0016; idem, *READing handwritten documents. Projekt READ und das Staatsarchiv Zürich auf dem Weg zur automatischen Erkennung von handschriftlichen Dokumenten*, *Geschichte und Informatik. Das Web als politische Herausforderung: historische Perspektiven*, Zürich 2019.

4 Srov. TRANSKRIBUS, *Ground Truth*, https://blog.transkribus.org/en/what-is-ground-truth.

5 Srov. TRANSKRIBUS, *Transkribus workflow*, https://www.transkribus.org/manuscripts.

6 Srov. TRANSKRIBUS, *Transkribus sites*, https://www.transkribus.org/sites.

uživatelsky přívětivého webového rozhraní.[7]

Charakteristika modelu *Latin Sermons 11th–12th C.*

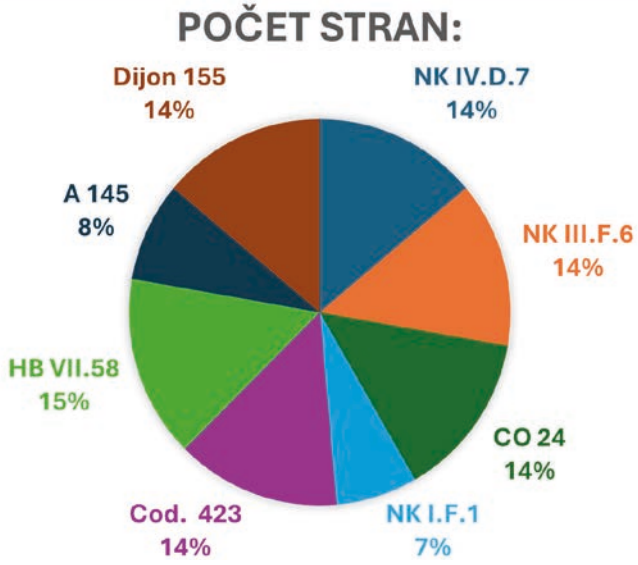
HTR model *Latin Sermons 11th–12th C.* vznikl jako součást bakalářské práce řešené na Katolické teologické fakultě Univerzity Karlovy v akademickém roce 2024/2025.[8] Cílem práce bylo vytvořit a popsat postup tvorby modelu pro automatické rozpoznávání rukopisných textů (*Handwritten Text Recognition*, dále HTR). Cílem nebylo jen vytvořit funkční nástroj pro automatizovaný přepis vybraných textů, ale rovněž důkladně zmapovat celý proces jeho tvorby a identifikovat faktory ovlivňující kvalitu a přesnost výsledného přepisu. Model byl koncipován jako specializovaný nástroj, optimalizovaný pro rozpoznávání latinských rukopisů psaných převážně pozdní karolinskou a raně gotickou minuskulou. Jeho aplikovatelnost je tedy omezena především na tento konkrétní typ písma. Samotný model zatím není veřejně dostupný. Tréninková data (*training data*), na jejichž základě byl model vytvořen, však budou v nejbližší době sdílena prostřednictvím repozitáře *Zenodo*.

Tréninková data (*training data*) připravená pro účely tvorby modelu *Latin Sermons 11th–12th C.* pocházejí z osmi rukopisů uložených v knihovnách pěti evropských zemí. Největší část tvoří rukopisy bohemikální provenience (srov. Graf 1). Časově pokrývají období 11. a 12. století a žánrově se soustředí na homiletické texty, které čerpají převážně z patristické tradice. Rukopisy byly záměrně vybrány s ohledem na čitelnost; většina z nich je psána pozdní karolinskou minuskulou, s výjimkou kodexu CO 24, jenž již reprezentuje raně gotickou minuskulu. Všechny texty jsou uspořádány v jednom textovém sloupci. Interlineární glosy a marginálie, pokud se v rukopise objevují, nebyly do tréninkových dat (*training data*) zahrnuty. Strany byly vybrány tak, aby poskytly reprezentativní vzorek pro trénování modelu *Latin Sermons 11th–12th C.* (srov. Graf 2). Tréninková folia pocházejí z různých částí jednotlivých kodexů, aby byla zajištěna dostatečná rozmanitost písařských rukou. Tréninková data tak zahrnují široké spektrum stylů písma, které se mohou v rámci jednoho rukopisu lišit.

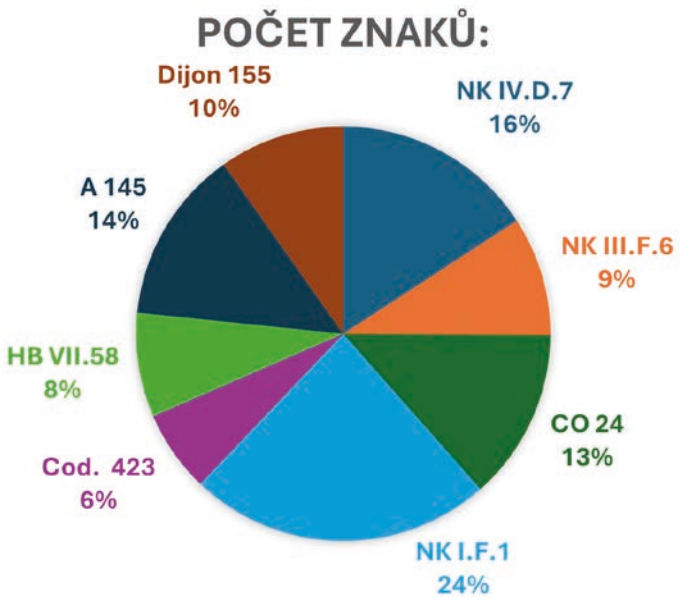
7 K problematice eScriptoria srov. Peter A. STOKES – Benjamin KIESSLING – Daniel STÖKL BEN EZRA a kol., *The eScriptorium VRE for Manuscript Cultures*, 2021, https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/. Ariane PINCHE, *Guide de transcription pour les manuscrits du Xe au XVe siècle*, 2022, https://hal.science/hal-03697382. Thibault CLÉRICE – Malamatenia VLACHOU-EFSTATHIOU – Alix CHAGUÉ, *CREMMA Medii Aevi: Literary manuscript text recognition in Latin, Journal of Open Humanities Data 9,* 2023, https://doi.org/10.5334/johd.97. Alžběta ZAVŘELOVÁ, *Projekt PERO – OCR pro historické texty, Duha: Informace o knihách a knihovnách 34,* 2025, č. 4, http://duha.mzk.cz/clanky/projekt-pero-ocr-pro-historicke-texty.

8 Barbora KULHOVÁ, *Technologie HTR a její využití pro přepis homilií z období 11.–12. století*, bakalářská práce, Katolická teologická fakulta, Univerzita Karlova, Praha 2025.

Graf 1: Výsečový graf zobrazuje zastoupení jednotlivých rukopisů v datasetu *Latin Sermons 11th–12th C.* v procentech s ohledem na počet stran



Graf 2: Výsečový graf zobrazuje zastoupení jednotlivých rukopisů v datasetu *Latin Sermons*



Analýza a výběr transkripčních zásad

Pro vytvoření modelu je klíčové připravit kvalitní tréninková data (*training data*), která zahrnují digitální kopie rukopisů a jejich konzistentní a bezchybné transkripce sloužící jako referenční texty. Takto sestavený soubor digitálních kopií a odpovídajících transkripcí se v odborné literatuře označuje jako *Ground Truth*. Přesnost a konzistence těchto transkripcí přímo ovlivňují kvalitu výsledného modelu. Již na počátku je proto nezbytné pečlivě zvážit vhodný transkripční přístup, a to s ohledem na konkrétní charakter daného rukopisu i zamýšlený cíl projektu.[9] Jako určující se nakonec ukázaly zejména dva faktory: kompatibilita s možným *base modelem* a míra uplatnitelnosti zvoleného přístupu na latinský rukopisný materiál z 11. a 12. století.

U transkripčních pravidel byly posuzovány způsoby řešení následujících problémů:

1. způsob přepisu dlouhého „ſ “,
2. způsob přepisu netečkované „ı“ a „i“,
3. způsob přepisu zdánlivého „-ij“,
4. rozlišování mezi „u“ a „v“,
5. problematika „e“, „ę“,
6. úpravy majuskulí na začátku vět, vlastních jmen a u obecných substantiv označujících konkrétní osoby (např.: Dominus, Salvator),
7. opravy nedůsledností v pravopise, záměrné změny hláskových podob slov, přidávání a ubírání písmen,
8. interpunkce,
9. spojování předložky se slovem,
10. číslovky,
11. zkratky.

Klíčovou práci na téma transkripčního přístupu v kontextu digitálních technologií představují zásady sestavené Peterem Robinsonem a Elizabeth Solopovou v projektu *Canterbury Tales Project*.[10] Jejich význam dnes spočívá hlavně v terminologii, kterou používá projekt *CATMuS*.[11] Rozdělují zde transkripci na čtyři stupně: *graphic, graphetic, graphemic* a *regularized*.

Graphic transcription věrně kopíruje všechny znaky v rukopise, včetně jejich délky, šířky a vzdáleností mezi jednotlivými znaky. Například v případě písmena „i“ je znak

9 Srov. TRANSKRIBUS, *What is Ground Truth*, https://readcoop.eu/what-is-ground-truth/.

10 Peter ROBINSON – Elizabeth SOLOPOVA, *Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue*, San Marino 1993, s. 19, https://doi.org/10.5281/zenodo.4050360.

11 Ariane PINCHE – Thibault CLÉRICE – Alix CHAGUÉ a kol., *CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts: A generalized set of guidelines and models for Latin scripts from Middle Ages (8th–16th century)*, 2023, https://catmus-guidelines.github.io/html/guidelines/en/generalites.html.

rozdělen na samostatný dřík a tečku a následně přepsán tak, aby přesně odpovídal tvaru a rozměru znaku. *Graphetic transcription* zachovává různé tvary písmen, kupříkladu u písmena „r" rozlišuje směřování a tvar příčky. Pojmem *graphemic transcription* nazýváme přepis, který zachovává původní zápis grafémů, tedy jednotlivých písmen či znaků. Oproti *graphetic transcription* nebere v potaz rozdílné tvary písmen, zaznamenává tedy pouze funkci znaku v abecedním systému. Posledním stupněm je *regularized transcription*, kdy je veškerý pravopis v rukopise upraven podle předem stanovené normy. Solopová s Robinsonem přiznávají, že v podstatě všechny dosud používané transkripční zásady nelze jednoznačně zařadit do jedné kategorie.[12] *Regularized transcription* nevolí z důvodu její náročnosti a také možnosti ztráty informací z primárního pramene, které mohou být důležité pro další výzkum. Taktéž vylučují užití *graphic transcription* vzhledem k charakteru jimi studovaného korpusu pramenů. Nakonec se přiklání k transkripci *graphemic transcription*.[13]

Alternativní terminologii a kategorizaci navrhují Estelle Guéville a David Joseph Wrisley. Ve své studii poukazují na výzvy a přínosy, které strojové učení (*machine learning*) přináší, a vše demonstrují na příkladu vlastního výzkumu latinských středověkých biblí. Transkripční přístupy dělí na: normalizovaný, semi-diplomatický a diplomatický.[14] Diplomatický přepis je v podstatě přepisem grafémickým. Semi-diplomatický přístup zachovává některé specifické tvary znaků, zkratky rozšiřuje, ale jejich rozšíření viditelně označuje například pomocí kurzívy nebo podtržení; zároveň se snaží věrně napodobit pravopis rukopisu.[15] Normalizovanou transkripci definují jako přepis textu upravený dle moderních standardů pravopisu a interpunkce tak, aby byl srozumitelný pro badatele a studenty. Zdůrazňují, že tento proces zjednodušení může vést ke ztrátě informací obsažených v původním rukopisu, které mohou být důležité pro některé typy výzkumu.[16] Guéville a Wrisley volí diplomatickou transkripci na základě účelu, pro který přepis vytvářejí.[17]

V kontextu vytváření transkripčních pravidel přizpůsobených automatickému rozpoznávání rukopisného textu (*Handwritten Text Recognition*, dále HTR) považujeme za vhodné představit diverzitu přístupů jako spektrum směřující od přepisu grafémického k přepisu regularizovanému či od přístupu diplomatického k normalizovanému. Výběr pravidel by měl respektovat účel tvorby automatického

12 Ibidem.
13 Ibidem.
14 Estelle GUÉVILLE – David Joseph WRISLEY, *Transcribing medieval manuscripts for machine learning*, *Journal of Data Mining and Digital Humanities*, 2022, https://doi.org/10.48550/arXiv.2207.07726.
15 Ibidem.
16 Ibidem.
17 Ibidem.

přepisu a zachovávat informace, které jsou později důležité pro výzkum.

Polovinu analyzovaného korpusu představují rukopisy bohemikální provenience. Na rukopisy tohoto typu se nejčastěji uplatňují ediční zásady Bohumila Ryby.[18] Tyto zásady se v zásadě řídí tím, že nemění hláskovou podobu slov, tedy nic nemění ani nepřidávají, pokud se nejedná o jevy pouze grafické povahy.[19] Jejich hlavním specifikem je rozlišování mezi „u" a „v". Tento způsob není pro zahraniční edice v dnešní době již běžný. Interpunkci v rukopise nahrazují interpunkcí logickou podle současného českého úzu. Velká počáteční písmena na začátku vět, vlastních jmen a u obecných substantiv označujících konkrétní osoby (např. *Dominus, Salvator*) považuje Ryba za jev čistě grafický a s ohledem na základní zásadu doporučuje v těchto případech psát velká písmena i v případě, kdy se v rukopise nevyskytují. Římské číslice transkribuje velkými římskými znaky i tehdy, pokud jsou v rukopise ve tvaru minuskulním. Výjimkou je vročení, u něhož doporučuje transkripci co nejvěrnější rukopisu (srov. Tab. 1).[20]

Tab. 1: Tabulka obsahuje transkripční pravidla Bohumila Ryby

Analyzované problémy	Bohumil Ryba
Dlouhé „ſ"	*Na „s"*
Netečkované „ı" a „i"	*Na „i"*
Zdánlivé „-ij"	*Na „-ii"*
Rozlišování „u" a „v"	*Rozlišuje „u" a „v"*
Majuskuly	*Upravuje*
Problematika „e", „ę"	*Dle rukopisu*
Opravy	*Ne*
Interpunkce	*Dle současného českého úzu*
Spojování předložek se slovem	*Nenapodobuje dle rukopisu*
Zkratky	*Rozepisuje*
Číslovky	*Upravuje minuskule na majuskule u římských číslovek, s výjimkou vročení*

S ohledem na rukopis *Homiliáře opatovického*, který byl edičně zpracován již třemi editory – Ferdinandem Hechtem, Konstantinem Miklíkem a Zdeňkem Uhlířem – jsme

18 Transkripční pravidla Bohumila Ryby byla původně určena pro vydání Husových spisů. Pravidla byla v omezené verzi vydána Centrem medievistických studií v roce 2000. Problematice dalších vydání se věnuje Jan ODSTRČILÍK, *Bohumil Ryba's Rules for Transcription of Latin Manuscript Texts (English Translation)*, 2024, https://doi.org/10.5281/zenodo.11060033. Jejich rozsáhlý výbor je také součástí paleografické čítanky: Dalibor HAVEL – Helena KRMÍČKOVÁ, *Paleografická čítanka: literární texty*, Brno 2014, s. 89–96.

19 Ibidem.

20 Ibidem.

se zaměřili také na ediční přístupy v těchto edicích.[21] V případě existujících edic nás primárně zajímala jejich využitelnost při tvorbě transkripcí v *Ground Truth*. Využitelnost edic Miklíka a Hechta limituje zejména absence kritického aparátu. V kontextu automatického přepisu totiž obecně platí zásada transkribovat text co nejvěrněji tak, jak je skutečně zachycen v rukopise.[22] Pokud edice nedisponuje informacemi o editorských zásazích, je nezbytné ji konfrontovat s primárním pramenem. Z tohoto důvodu jsme se rozhodli použít stávající edice pouze jako doplňkový referenční rámec, nikoli jako přímý zdroj tréninkových dat (*training data*). Transkripční pravidla proto nebyla přizpůsobována těm, která byla uplatněna v daných edicích. Vedle existujících edic a české transkripční normy jsme se dále zaměřili na přístupy používané přímo v kontextu automatického rozpoznávání textu (*automatic text recognition*). Za grafémické označujeme transkripční zásady, které v roce 2022 vypracovala Ariane Pinche.[23] Ty byly dále rozšířeny v roce 2023.[24] Transkripční pravidla, která vznikla syntézou zásad z let 2022 a 2023, jsou dnes zveřejněna jako zásady *CATMuS* (*Consistent Approaches to Transcribing ManuScripts*). Jsou navržena tak, aby byla kompatibilní s rukopisy psanými v latinském, španělském, italském a francouzském jazyce v období 10. – 15. století, a zároveň jsou uzpůsobena tak, aby podporovala strojové učení (*machine learning*) (srov. Tab. 2). [25] Pinche ve svém přístupu upouští od rozepisování zkratek. Argumentuje tím, že rozšiřování zkratek vede ke ztrátě informací obsažených v původním prameni. Samotný akt rozšíření představuje interpretační zásah, který nutně vzdaluje transkripci od grafické podoby rukopisu. Zároveň není možné stanovit jednotná pravidla pro rozepisování zkratek v rukopisech různých jazyků a období. Zachovávání zkratek považuje za optimální přístup jak pro trénování nových, tak i pro rozšiřování starších modelů. Středověké zkratky jsou v tomto rámci reprezentovány pomocí znaků definovaných standardem *Medieval Unicode Font Initiative* (*MUFI*).[26]

21 Ferdinand HECHT, *Beiträge zur Geschichte Böhmens. Abtheilung I. I. Band. Saec. XII: Quellensammlung. Das Homiliar des Bischofs von Prag*, Praha 1863. Josef MIKLÍK, *Opatovský homiliář*, *Časopis katolického duchovenstva* 72, 1931, č. 1–10; Josef MIKLÍK, *Opatovský homiliář*, *Časopis katolického duchovenstva* 73, 1932, č. 11–10. Ediční práci na Homiliáři opatovickém se věnoval také Zdeněk UHLÍŘ, jehož edice byla spolu s digitalizovanou verzí zveřejněna v roce 2010 v databázi *Manuscriptorium*. Uhlířova práce však není v současnosti dostupná.

22 Estelle GUÉVILLE – David Joseph WRISLEY, *Transcribing medieval manuscripts for machine learning*, *Journal of Data Mining and Digital Humanities*, 2022, https://arxiv.org/abs/2207.07726.

23 Ariane PINCHE, *Guide de transcription pour les manuscripts du Xe au XVe siècle*, 2022, https://hal.science/hal-03697382.

24 Thibault CLÉRICE – Malamatenia VLACHOU-EFSTATHIOU – Alix CHAGUÉ, *CREMMA Medii Aevi: Literary manuscript text recognition in Latin*, *Journal of Open Humanities Data 9*, 2023, https://doi.org/10.5334/johd.97.

25 A. PINCHE – T. CLÉRICE – A. CHAGUÉ a kol., *CATMuS-Medieval*.

26 A. PINCHE, *Guide de transcription*. Medieval Unicode Font Initiative (MUFI), https://mufi.info/.

Tab. 2: Tabulka obsahuje transkripční pravidla CATMuS

Analyzované problémy	CATMuS
Dlouhé „ſ“	*Na„s“*
Netečkované „ı“ a „i“	*Na„i“*
Zdánlivé „-ij“	*Na„-ii“*
Rozlišování mezi „u“ a „v“	*Nerozlišuje*
Majuskule	*Dle rukopisu*
„e“, „ę“	*Dle rukopisu*
Opravy	*Ne*
Interpunkce	*Grafémická a zjednodušená*
Spojování předložek se slovem	*Dle rukopisu*
Zkratky	*Zachovány*
Číslovky	*Dle rukopisu*

Vzhledem k povaze našich tréninkových dat (*training data*), která zahrnují výhradně latinské rukopisy, se pravidla *CATMuS* neukázala jako prakticky využitelná. Tato pravidla, vytvořená pro různé jazyky a historická období, předpokládají nahrazování středověkých zkratek speciálními znaky podle standardu *Medieval Unicode Font Initiative* (*MUFI*), což by bylo nejen časově náročné, ale zároveň i v rozporu s cílem vytvořit čtenářsky přístupnou transkripci.

Při analýze transkripčních zásad bylo přihlédnuto také ke způsobu transkripce uplatňovanému během *HTR Winter School 2023/2024*, pořádané *Österreichische Akademie der Wissenschaften*.[27] Jedna ze skupin pod vedením Tima Geelhaara vytvořila přepis několika folií *Liber Pontificalis*, zatímco druhá skupina pod vedením Jana Odstrčilíka se věnovala pozdně středověkému rukopisnému materiálu (srov. Tab. 3).[28] První skupina přitom aplikovala stejná pravidla, která byla původně formulována Geelhaar(em) při tvorbě modelu *Carolingian Minuscule Model (CMM) 9th–11th C.* S ohledem na plánované využití tohoto modelu jako *base modelu* jsme zvolili stejný transkripční přístup i pro přípravu našich tréninkových dat (*training data*) při trénování modelu *Latin Sermons 11th–12th C.* Konkrétní podoba těchto

27 Srov. Theresa HALBERTSCHLAGER – Cinzia GRIFONI – Leon PÜRSTINGER – Evina STEIN – Alice MORANDY – Evangelos ADAM – Tatiana TOMMASI – Essi NUUTINEN – Veronika WLADIKA – Richard M. POLLARD – William WEIS, *HTR Winter School 2023/2024 – Carolingian Latin* [Dataset], *Zenodo*, 2024, https://doi.org/10.5281/zenodo.10610704.

28 Srov. Jan ODSTRČILÍK – Fatih YÜCEL – Liene ROKPELNE – Ana MIHALJEVIĆ – Zuzana ŠALDOVÁ – Adéla ZELENKOVÁ – Mara CIUNTU – Magdaléna LUKÁČ KOVÁČOVÁ – Andrej VAŠÍČEK – Viktória CEHUĽOVÁ – Ivana LUKÁČ LABANCOVÁ – Leonhard ENGELMAIER – Andrea SCALIA – Albert KOHN, *HTR Winter School 2023/2024 – Late Medieval Latin, ONB 3891 (1.0.0)* [Dataset], *Zenodo*, 2024, https://doi.org/10.5281/zenodo.10589479.

transkripčních zásad je podrobně rozpracována v následující kapitole.[29]

Tab. 3: Tabulka obsahuje transkripční pravidla vytvořené během HTR Winter School 2023/2024

Analyzované problémy	Jan Odstrčilík	Tim Geelhaar
Dlouhé „ſ“	*Na „s“*	*Na „s“*
Netečkované „ı“ a „í“	*Na „i“*	*Na „i“*
Zdánlivé „-ij“	*Na „-ii“*	*Na „-ii“*
Rozlišování mezi „u“ a „v“	*Rozlišuje „u“ a „v“*	*Nerozlišuje*
Majuskule	*Dle rukopisu*	*Dle rukopisu*
„e“, „ę“	*Dle rukopisu*	*Dle rukopisu*
Opravy	*Ne*	*Ne*
Interpunkce	*Dle rukopisu*	*Dle rukopisu*
Spojení předložky se slovem	*Nenapodobuje dle rukopisu*	*Nenapodobuje dle rukopisu*
Zkratky	*Rozepisuje*	*Rozepisuje*
Číslovky	*Dle rukopisu*	*Dle rukopisu*

Transkripční zásady pro model *Latin sermons 11th–12thC.*

Transkripční přístup následuje zásady vytvořené Timem Geelhaarem pro model *Carolingian Minuscule Model (CMM) 9th–11th C.*[30] Zásady jsou rozděleny do tří kategorií: rozložení, znaky a interpunkce. Rozložení zahrnuje pravidla pro zachování struktury textu, jako je rozmístění řádků a pravidla pro práci s iniciálami, margináliemi a interlineárními glosami. Kategorie „znaky“ se zaměřuje na přesné zaznamenání jednotlivých grafémů, včetně variant tvarů písmen, ligatur a zkratek. Kategorie „interpunkce“ se věnuje způsobu přepisu interpunkčních znamének a jejich simplifikaci.

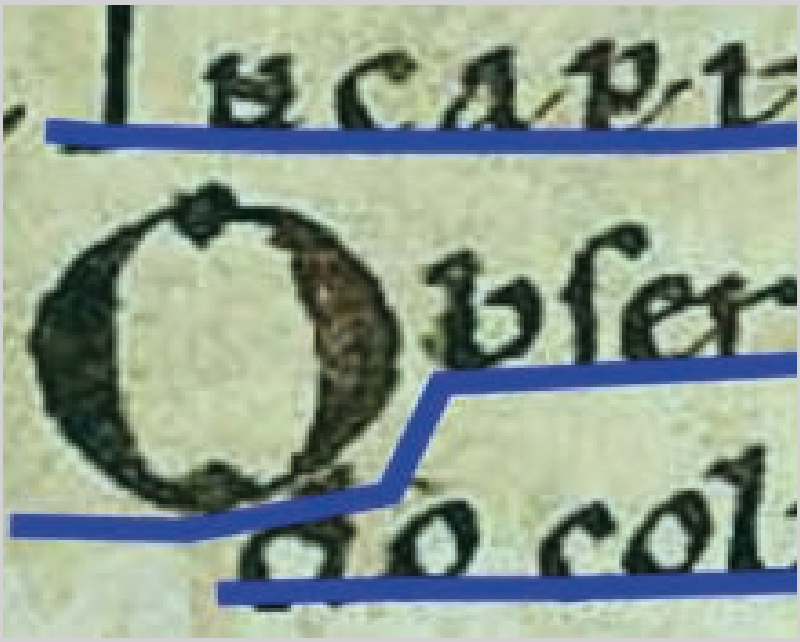
Rozložení

Přepisován je pouze hlavní text bez interlineárních glos, margináliı, korekcí a číslování stran. Pozdější přípisky nejsou brány v potaz. Iniciály jsou zařazeny do hlavního textu pouze tehdy, pokud svou velikostí nepřesahují více než dva řádky. Pokud iniciála nepřesahuje svou velikostí dva řádky, je začleněna do hlavního textu tak, že je jí manuálně přizpůsoben řádek (srov. Obr. 1). Pokud je iniciála větší, je vynechána (srov. Obr. 2).

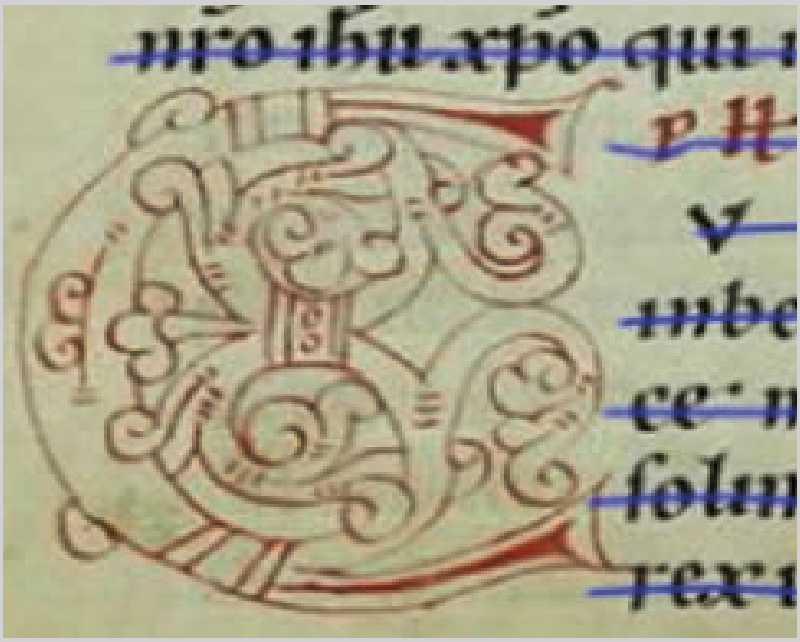
Špatně čitelné části jsou označeny pomocí značky (*tag*) „unclear“, zatímco porušené části rukopisu značkou (*tag*) „gap“; případně je porušení přizpůsoben řádek (srov. Obr. 3). Takto označené části jsou následně automaticky vynechávány během procesu trénování.

29 TRANSKRIBUS, *Latin–carolingian minuscule*, https://readcoop.eu/model/latin-carolingian-minuscule/.

30 TRANSKRIBUS, *Latin–carolingian minuscule*, https://readcoop.eu/model/latin-carolingian-minuscule/.



Obr. 1: Obrázek ukazuje způsob manuálního přizpůsobení řádku iniciále, která nepřesahuje dva řádky.



Obr. 2: Obrázek ukazuje způsob vynechání iniciály, která přesahuje více než dva řádky.

Obr. 3: Obrázek ukazuje způsob manuálního přizpůsobení řádků v případě poškození rukopisu.

Předchází se tak tomu, aby se model učil na nepřesných nebo neúplných datech, což by mohlo negativně ovlivnit jeho kvalitu.

Znaky

1. způsob přepisu dlouhého „ſ“,
2. způsob přepisu netečkované „ı“ a „i“,
3. způsob přepisu zdánlivého „-ij“,
4. rozlišování mezi „u“ a „v“,
5. problematika „e“, „ę“,
6. úpravy majuskulí na začátku vět, vlastních jmen a u obecných substantiv označujících konkrétní osoby (např.: Dominus, Salvator),
7. opravy nedůsledností v pravopise, záměrné změny hláskových podob slov, přidávání a ubírání písmen,
8. interpunkce,
9. spojování předložky se slovem,
10. číslovky,
11. zkratky.

Interpunkce je zjednodušena na znaky „?“ (003F), „·“ (00B7) a „;“ (003B). U teček není

rozlišena výška jejich položení a všechny jsou transkribovány s pomocí výše položené tečky „·“ (00B7).[31] V případě všech dvojitých interpunkčních znaků je použit středník (srov. Tab. 4).

Tab. 4: Tabulka demonstruje příklady transkripce interpunkce

Interpunkce v rukopisu	Způsob transkripce
	·
	·
	?
	;
	;

Manuální transkripce a analýza rozložení textu (*layout recognition*)

Transkribus doporučuje pro trénování modelu použít přibližně 25–75 stran obsahujících 5 000–15 000 slov.[32] Naše tréninková data zahrnují 64 stran, přičemž v některých případech se jedná o dvoustrany. Celkový rozsah činí 19 705 slov. U všech rukopisů byla k dispozici jejich digitalizovaná kopie, která byla nahrána do aplikace *Transkribus* ve formátu JPEG nebo PNG. Všechny strany byly manuálně transkribovány. Pro analýzu rozložení (*layout recognition*) jednotlivých folií byl použit model *Universal Lines*[33] určený k automatickému rozpoznávání struktury strany (*layout recognition*). Jeho hlavní funkcí je detekce základních prvků stránky, jako jsou textové sloupce, odstavce, řádky, slova či obrazové segmenty.[34] Model *Universal Lines* je nejuniverzálnějším modelem na platformě *Transkribus* a měl by vyhovovat většině standardních rukopisů a tisků.[35] Výsledná automatická analýza rozložení (*automatic layout recognition*) textu na stránce byla následně opravena na základě transkripčních pravidel pro rozložení. Zachován byl výhradně hlavní textový blok. Zvláštní pozornost byla věnována iniciálám, které byly

31 Interpunkční znaky doplňuji o Unicode kódy kvůli jednoznačné typografické identifikaci znaků, srov. *UNICODE – Latin-1 Supplement*, https://jrgraphix.net/r/Unicode/

32 TRANSKRIBUS, *Model Setup and Training*, https://help.transkribus.org/model-setup-and-training.

33 Ibid. [cit. 6. 11. 2024].

34 TRANSKRIBUS, *Automatic layout recognition*, https://help.transkribus.org/automatic-layout-recognition.

35 TRANSKRIBUS, *Baselines models*, https://help.transkribus.org/baselines-models.

do hlavního textu zařazeny pouze v případě, že svou velikostí nepřesahovaly dva řádky. V takových případech byly příslušné řádky manuálně upraveny, aby iniciála odpovídala lineárnímu toku textu. Ve většině případů nebyla nutná další manuální úprava řádků. Pouze ojediněle docházelo k problémům s nesprávným rozdělením řádků, což bylo pravděpodobně způsobeno nižší kvalitou jedné z digitálních kopií. Tento konkrétní rukopis se při prvním stažení nepodařilo uložit v plném rozlišení, a byl proto znovu nahrán ve vyšší kvalitě. V několika případech bylo nezbytné upravit řádky, které nedosahovaly až ke konci textového sloupce. Poškození rukopisů byla ojedinělá a nepředstavovala pro automatickou analýzu rozložení závažnější komplikace.

Využití *base modelu*

Jednou z klíčových funkcí nabízených platformou *Transkribus* při tvorbě nového HTR modelu je možnost využití *base modelu*. Tento postup umožňuje algoritmu přenést znalosti z již existujícího modelu do nově vytvářeného modelu, čímž může výrazně zefektivnit proces trénování a zároveň zlepšit přesnost výsledného přepisu. Předem vytrénovaný model, který byl vytvořen na základě obdobných rukopisů, v podstatě rozšiřuje objem tréninkových dat (*training data*). Tento přístup je zvláště vhodný v případech, kdy badatel nedisponuje dostatečně rozsáhlým vlastním tréninkovým korpusem.[36]

Výběr vhodného *base modelu* je ovlivněn několika faktory. Mezi nejdůležitější patří jazyk textu, jeho datace a typ písma, které by měly co nejlépe odpovídat charakteru analyzovaných rukopisů. V tomto projektu byl jako výchozí model zvolen *Carolingian Minuscule Model (CMM) 9th–11th C.* vytvořený Timem Geelhaarem. Tento model byl vytrénován na základě 46 různých rukopisů pocházejících z období let 800–1100, psaných karolinskou minuskulou. Současným kurátorem modelu je Leon Pürstinger (*Österreichische Akademie der Wissenschaften*). Protože se časové rozmezí modelu kompletně nekryje s datací rukopisů použitých v našem výzkumu, rozhodli jsme se jej otestovat na několika foliích. Výsledky prokázaly dostatečnou přesnost přepisu, což potvrdilo jeho využitelnost jako *base modelu* pro náš tréninkový proces.

Trénování modelu

Vzniku finálního modelu předcházelo několik modelů testovacích, u nichž byl sledován vliv klíčových faktorů, jako je množství tréninkových dat (*training data*), počet tréninkových cyklů (*epochs*) a použití *base modelu*. Cílem těchto testů bylo optimalizovat uvedené proměnné pro model *Latin Sermons 11th–12th C.* Všechny testy uvedené v tabulce byly vyhodnoceny pomocí stejného validačního setu

36 Béatrice COUTURE – Ferah VERRET – Maxime GOHIER – Dominique DESLANDERS, *The Challenges of HTR Model Training: Feedback from the Project*, Montreal 2023.

(*validation set*) (srov. Tab. 5), přičemž bylo sledováno procento chybovosti přepisu (*CER – Character Error Rate*).[37]

Tab. 5: Tabulka slouží k porovnání hodnot mezi testovacími modely, porovnávány jsou počty stran tedy množství tréninkových dat *(training data)*, využití base modelu (BM), hodnota CER na validačním setu (VS) a počet tréninkových cyklů (epochs)

	Test 1	**Test 2**	**Test 3**	**Test 4**
Počet stran	*24*	*55*	*55*	*55*
BM ano/ne	*ANO*	*ANO*	*ANO*	*NE*
CER na VS	*15,46%*	*7,81%*	*7,95%*	*8,26%*
Reálný počet epoch	*100*	*100*	*150*	*100*

První test ukázal, že počet 24 stran tréninkových dat (*training data*) nepředstavoval dostatečný objem pro dosažení uspokojivých výsledků. Model vykazoval vysokou hodnotu *CER* (15,46 %) na validačním setu (*validation set*). V následujících testech (označených jako testy 2, 3 a 4) byla pozornost zaměřena na optimalizaci počtu tréninkových cyklů (*epochs*) a vliv použití *base modelu*. Testy 2 a 3, u nichž byl již navýšen počet tréninkových dat (*training data*), přinesly výrazně lepší výsledky ve srovnání s testem prvním. Hodnoty *CER* byly u těchto dvou testů téměř totožné a lišily se pouze o několik desetin procenta. Tento rozdíl lze částečně přičíst rozdílnému počtu tréninkových cyklů (*epochs*), pravděpodobněji však odráží drobné odchylky vznikající během samotného procesu trénování. Trénování neuronových sítí zahrnuje určitou míru náhodnosti, a tudíž může docházet k mírným odchylkám i tehdy, jsou-li modely trénovány za stejných podmínek. Porovnání výsledků testů 2 a 3 nás přivedlo k závěru, že použití 100 tréninkových cyklů (*epochs*) představuje pro trénování modelu optimální hodnotu. Test 4, ve kterém nebyl použit *base model*, ukázal mírné zvýšení hodnoty *CER* ve srovnání s testy 2 a 3. Tento výsledek potvrzuje, že využití *base modelu* významně přispívá ke zlepšení výkonu a snižuje chybovost přepisu. Na základě provedených testů lze konstatovat, že pro dosažení co nejnižší hodnoty *CER* na validačním setu (*validation set*) je v našem případě zásadní kombinace tří faktorů: dostatečného množství tréninkových dat (*training data*), využití vhodného *base modelu* a nastavení optimálního počtu 100 tréninkových cyklů (*epochs*).

Hodnocení kvality modelu na základě hodnoty CER na validačním setu

37 CER (*Character Error Rate*), který vyjadřuje míru chybovosti přepisu. Hodnota CER se vypočítává dle vzorce: CER = [(i + s + d) / n] × 100, kde *n* představuje celkový počet znaků, *i* počet nesprávných vložení, *s* počet nesprávných nahrazení a *d* počet vymazaných znaků. Hodnota CER pod 10 % je obecně považována za přijatelnou.

Kvalita výsledného modelu je hodnocena na základě hodnoty *CER* (*Character Error Rate*) na validačním setu (*validation set*). Validační set je dataset, který nebyl použit při trénování modelu. Platforma *Transkribus* nabízí dva způsoby jeho výběru: manuální, kdy uživatel sám označí konkrétní strany, a automatický, při kterém algoritmus vyčlení určité procento materiálu z tréninkového datasetu. Pro evaluaci modelu byly využity dva validační sety (*validation sets*). Validační set A byl vytvořen přímo v prostředí *Transkribus*. Během tréninkového procesu byly vybrané strany označeny jako validační data (*validation data*), tzn. nebyly použity pro učení a sloužily výhradně k hodnocení výkonnosti modelu. Tento set obsahuje přibližně 17 000 znaků rozdělených do osmi stran, přičemž z každého rukopisu obsaženého v tréninkovém datasetu byla vybrána právě jedna strana. Naproti tomu validační set B byl vytvořen mimo platformu *Transkribus*. Obsahuje další sadu osmi stran, které model nikdy neviděl ani při trénování, ani během validace. Také v tomto případě pochází každá strana z jiného rukopisu obsaženého v tréninkovém datasetu a snaží se co nejvěrněji odpovídat charakteru setu A (písařská ruka, typ textu, výzvy pro model – drobné poškození, větší iniciála). V praxi to znamená, že hodnota *CER* na validačním setu A byla získána přímo v aplikaci *Transkribus* po dokončení trénování modelu. Náš model dosáhl hodnoty *CER* 5,7 % na validačním setu A, což představuje poměrně vysokou míru přesnosti. Naproti tomu model *Carolingian Minuscule Model (CMM) 9th–11th C.*, který jsme použili jako *base model*, vykázal na stejném validačním setu hodnotu *CER* 13,7 %.

Rozdíl téměř osmi procentních bodů jednoznačně ukazuje, že vlastní trénink přinesl výrazné zlepšení výsledků. Hodnota *CER* externího validačního setu B byla získána na základě manuálního přepisu dalších osmi stran, z nichž každá pocházela z jiného rukopisu obsaženého v tréninkovém datasetu. Přepis těchto stran sloužil jako referenční text, který byl následně porovnán s automatickým přepisem vytvořeným modelem *Latin Sermons 11th–12th c.* v prostředí aplikace *Transkribus*. Při vyhodnocování přesnosti byly sledovány dvě varianty chybovosti, označené jako B1 a B2. V prvním případě *CER VS B (B1)* byly za chyby považovány všechny nesprávně rozpoznané části textu, včetně interlineárních glos a marginálií. Druhá hodnota *CER VS B (B2)* zohledňuje pouze chyby, kterých se model dopustil v hlavním textovém bloku, zatímco glosy a marginálie byly při výpočtu vynechány (srov. Tab. 6).

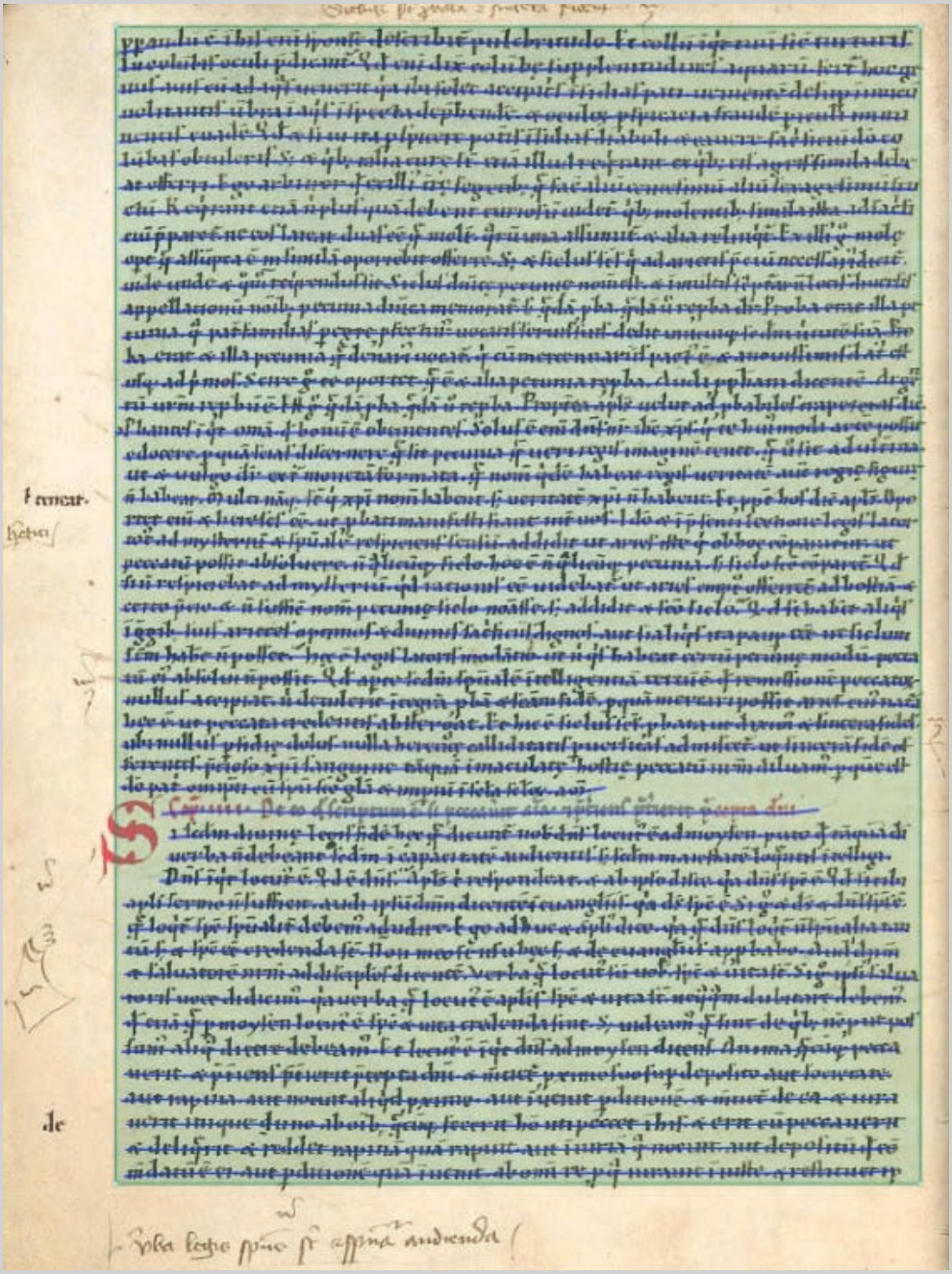
Tab. 6: Tabulka porovnává hodnoty CER na validačních setech (VS), VS A reprezentuje validační set vytvořený skrze platformu Transkribus, VS B (B1) a VS B (B2) reprezentují externí validační sety

	VS A	**VS B (B^1)**	**VS B (B^2)**
Hodnota CER	*5,7 %*	*16,42 %*	*3,17 %*

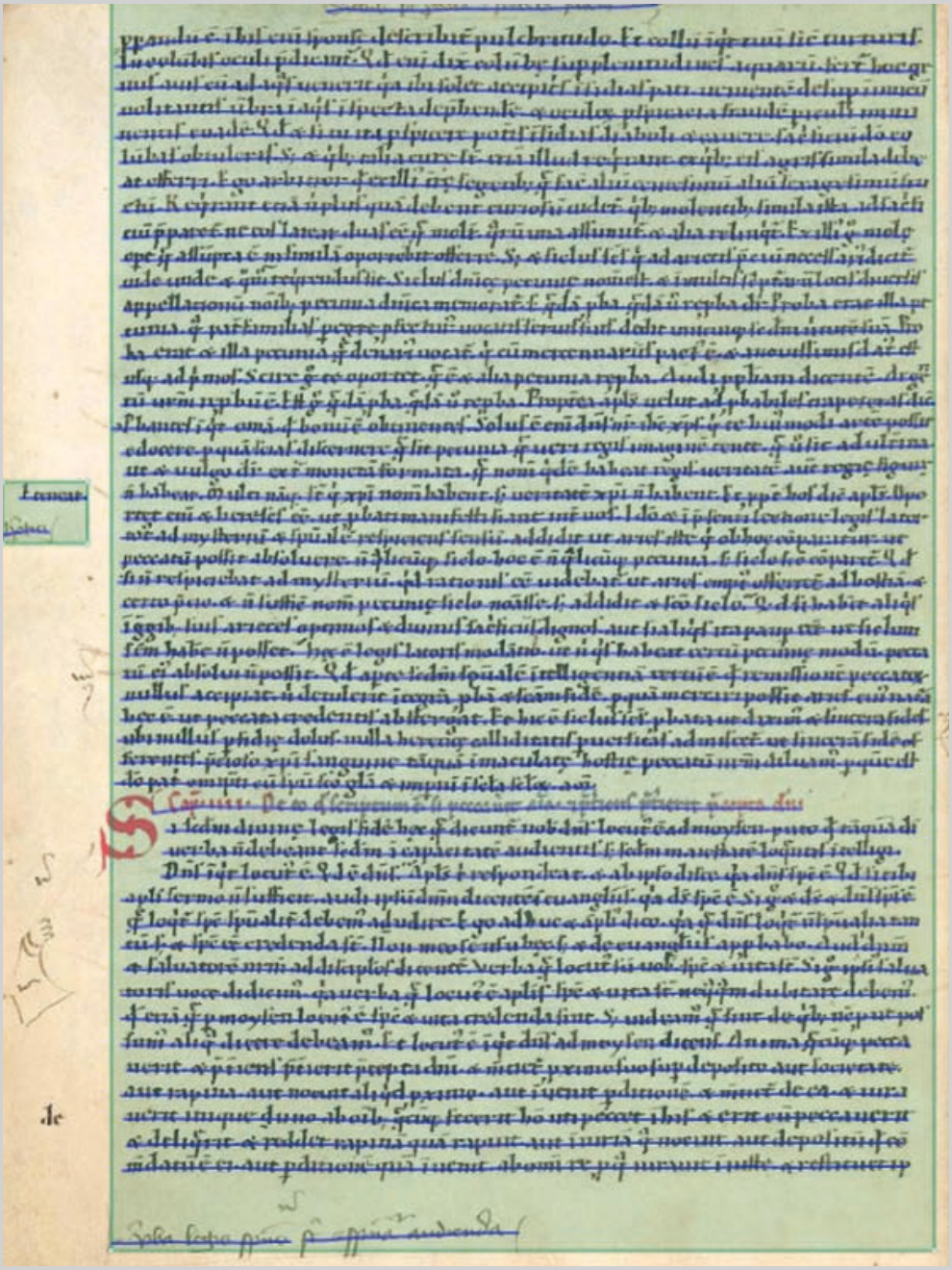
K výpočtu hodnot *CER* pro validační set B byl využit nástroj *CERberus*, který oproti standardnímu výstupu z *Transkribusu* poskytuje podrobnější analýzu chyb.[38] Tento nástroj nepracuje pouze s celkovou hodnotou *CER*, ale rozlišuje rovněž jednotlivé typy chyb, tedy počet znaků, které byly chybně nahrazeny, vymazány či nesprávně vloženy oproti referenčnímu přepisu. *CERberus* navíc poskytuje statistiku výskytu chyb podle jednotlivých znaků, například informaci o tom, kolikrát a v jakém procentu model chyboval při rozpoznávání konkrétního písmene. Tyto údaje umožňují lépe porozumět povaze chyb a představují důležitý nástroj pro cílenou optimalizaci tréninkového procesu. Test potvrdil, že model dosahuje výrazně lepších výsledků při přepisu pouze hlavního textového bloku (B2) než při zpracování celého folia včetně interlineárních glos a marginálií (B1). Porovnání ukazuje, že hlavní problém nespočíval ve schopnosti samotného automatického rozpoznávání písma (*Handwritten Text Recognition*, dále HTR), ale v selhávající segmentaci (*automatic layout recognition*), která chybně označovala části stránky, jež nebyly součástí tréninkového datasetu. Model tedy nebyl schopen respektovat pravidla ohledně rozložení strany a konzistentně rozlišovat mezi hlavním textem a dalšími prvky jejího rozvržení.

Z toho důvodu byl vytrénován nový model pro automatické rozpoznávání rozložení stránky (*automatic layout recognition model*). Při jeho tvorbě byla využita stejná tréninková data (*training data*), přičemž rozložení stran bylo již během jejich přípravy upraveno podle předem definovaných pravidel segmentace (*automatic layout recognition*). Cílem bylo vytvořit model, který bude jako textové pole označovat výhradně hlavní text a zároveň bude ignorovat interlineární glosy, marginálie, číslování stran a iniciály přesahující dva řádky. Výsledný *Latin Sermons 11th–12th C. baseline model* dosáhl na validačním setu (*validation set*) hodnoty *CER* 8,57 %. Srovnání schopností modelů *Universal Lines* a *Latin Sermons 11th–12th C. baseline model* je demonstrováno na ukázce dvou folií z rukopisů NK I.F.1 a CO 24 (srov. Obr. 4–7). Záměrně byla vybrána folia obsahující pozdější přípisky, marginálie a větší iniciály, které představují typické výzvy při segmentaci rukopisného materiálu.

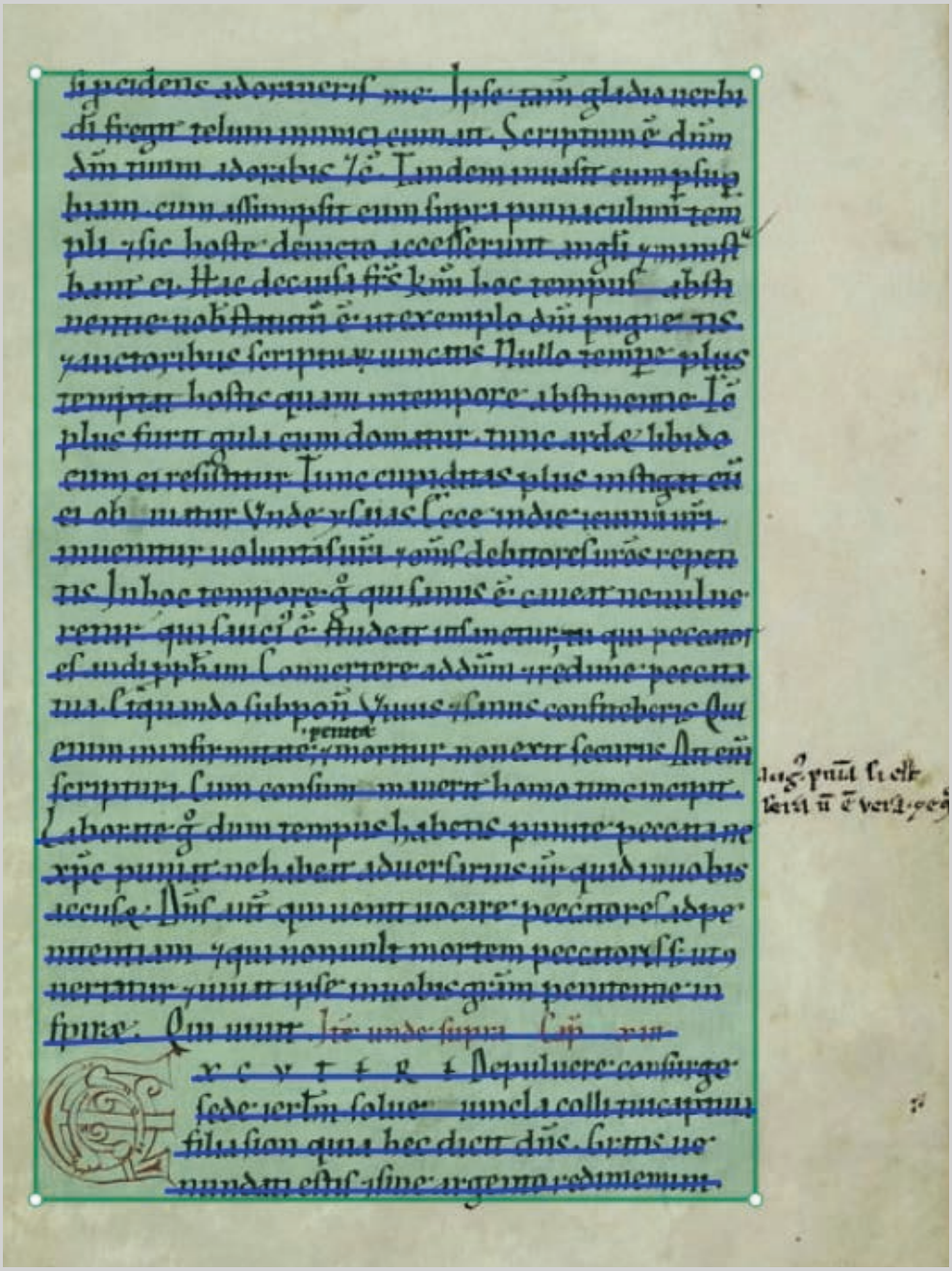
38 CERberus je jednoduchý nástroj pro digital humanities, který umožňuje generovat hodnotu CER. Do aplikace uživatel vloží referenční text a automatický přepis, tyto strany jsou porovnány a je na nich vypočtena hodnota CER. Srov. Wouter HAVERALS, *CERberus: Guardian Against Character Errors*, 2023, https://github.com/WHaverals/CERberus.



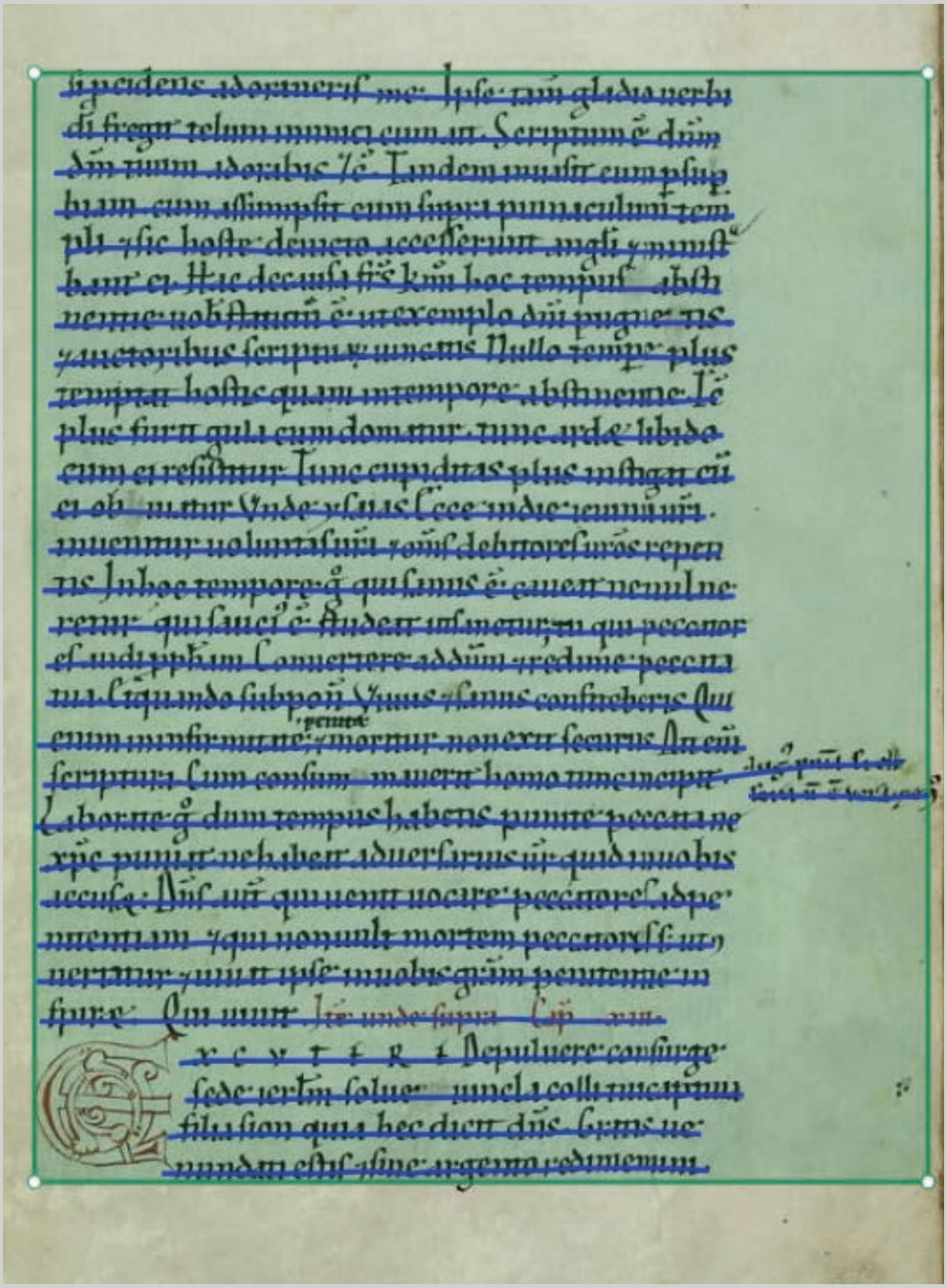
Obr. 4: Obrázek obsahuje analýzu rozložení provedenou modelem Latin sermons 11th–12th C. baseline model na foliu z rukopisu NK I.F.1.



Obr. 5: Obrázek obsahuje analýzu rozložení provedenou modelem Universial Lines na folium z rukopisu NK I.F.1.



Obr. 6: Obrázek obsahuje analýzu rozložení provedenou modelem Latin sermons 11th–12th c. baseline model na foliu z rukopisu CO 24.



Obr. 7: Obrázek obsahuje analýzu rozložení provedenou modelem Universal Lines na foliu z rukopisu CO 24.

Evaluace výsledků

Na demonstračních ukázkách je patrné, že *Latin Sermons 11th–12th C. baseline model* dokázal naplnit požadavky na automatické rozpoznávání rozložení stránky (*automatic layout recognition*). Maximální optimalizace výstupní transkripce tak může být dosažena kombinovaným použitím modelu pro rozložení strany (*automatic layout recognition model*) a modelu pro samotnou automatickou transkripci (*automatic text recognition model*). V praxi byla nejprve na demonstračních foliích provedena segmentace s pomocí *Latin Sermons 11th–12th C. baseline modelu*, načež byl stejný materiál automaticky přepsán HTR modelem *Latin Sermons 11th–12th C.*

Výsledná hodnota *CER* na demonstračním foliu z rukopisu NK I.F.1 činila 4,79 % a na foliu z rukopisu CO 24 dosáhla 10,39 %. V příloze je pro demonstraci přiložena automatická transkripce těchto folií. V rámci evaluace jsme se rovněž zaměřili na chybovost v rozpoznávání konkrétních znaků. Nejčastěji se chyby vyskytovaly u znaků, které neobvykle přesahují standardní řádek. Příkladem je znak „ę", zejména v případech, kdy kauda těsně nepřiléhá k písmenu. Podobný typ problému se vyskytoval u interpunkce, konkrétně u záměny středníku a tečky. Model často nedokázal správně zaznamenat celý znak a interpretoval středník jako výše položenou tečku. Další časté chyby se vyskytovaly u těsně umístěných liter s dříky nebo oblouky, jako jsou „m" a „n", případně „u" a „i". Docházelo tak k záměně „q" a „g", „a" za „e" a „e" za „c", což mohlo být zapříčiněno podobností těchto znaků v určitých kodexech. Překvapivě model zvládal relativně dobře rozšiřování zkratek, ačkoliv drobné chyby se objevily u koncovek typu „-us", „-um", „-ur", „-ar", případně u zkratky pro „-con/m". Problematické bylo také písmeno „z", které se v latinských textech vyskytuje relativně zřídka, což pravděpodobně přispělo k vyšší chybovosti při jeho přepisu. Model také občas chyboval při správném rozdělování předložky se slovem.

Přes uvedené nedostatky je třeba zdůraznit, že model vykazuje značný potenciál pro urychlení a zefektivnění výzkumné práce. Mnohé z chyb jsou marginálního charakteru a celková přesnost modelu je velmi vysoká. Dokládá to i dosažená hodnota *CER* 5,7 %, která potvrzuje spolehlivost modelu a jeho praktickou využitelnost. Platformu *Transkribus*, na které byl model vyvíjen a testován, lze hodnotit jako vysoce efektivní nástroj pro práci s automatickou transkripcí rukopisného materiálu. Nabízí intuitivní uživatelské rozhraní, které umožňuje snadnou správu projektů, trénování modelů a efektivní export dat i pro uživatele s omezenými technickými schopnostmi. Určité limity lze shledat v oblasti hodnocení kvality modelu, kde by identifikaci problémových oblastí usnadnila větší transparentnost. Platforma poskytuje širokou nabídku již existujících modelů a díky vysoké rychlosti zpracování dat a flexibilitě nástroje bylo možné dosáhnout vysoké efektivity při práci.

Autorky | Authors
Michala Falátková
Katolická teologická fakulta
Univerzita Karlova
Thákurova 3
160 00 Praha 6
michaela.falatkova@ktf.cuni.cz
Česká republika
ORCID: 0000-0002-5750-7262

Barbora Kulhová
Katolická teologická fakulta
Univerzita Karlova
Thákurova 3
160 00 Praha 6
Česká republika
barbora.kulhova@ktf.cuni.cz
ORCID: 0009-0006-6233-7268

Mgr. Michala Falátková, Ph.D. (* 1988) působí na Katedře církevních dějin a literární historie Katolické teologické fakulty Univerzity Karlovy a na Katedře pomocných věd historických a archivnictví Filozofické fakulty Univerzity Hradec Králové. Ve svých výzkumných aktivitách se zaměřuje na propojení klasických metod historického bádání s moderními přístupy digital humanities. V posledních letech zkoumá využití technologií HTR pro práci s latinskými středověkými rukopisy a věnuje se metodologickým otázkám zpřístupňování středověkých textů v digitálním prostředí.

Bc. Barbora Kulhová (* 2002) je studentkou navazujícího magisterského programu Dějiny evropské kultury na Katolické teologické fakultě Univerzity Karlovy. Ve svém výzkumu se zaměřuje na možnosti využití technologie HTR pro automatický přepis karolinských rukopisů. Při své práci používá především platformu Transkribus, která umožňuje trénink a nasazení modelů pro strojové rozpoznávání historického písma.

Historia Aperta published by Philosophical Faculty University of Hradec Králové.
ISSN 2788-0702 (Print) ISSN 2788-0710 (Online)
Platinum / Diamond Open Access model.

Summary

This study presents the development of a Handwritten Text Recognition (HTR) model tailored for Latin sermons from the 11th and 12th centuries, written in late Carolingian and early Gothic minuscule. The model, titled Latin Sermons 11th–12th C., was trained using data from eight manuscripts across five European countries. Emphasis was placed on analyzing transcription approaches and optimizing key factors such as training data volume, number of epochs, and the use of a base model. The final model achieved a Character Error Rate (CER) of 5.7 %, demonstrating strong adaptability to various scribal hands. The article also discusses the selection and justification of transcription guidelines, comparing established approaches (e.g., graphemic vs. regularized transcription) and their compatibility with machine learning processes. Manual transcription, layout recognition, and model training were carried out using the Transkribus platform. Additionally, the effectiveness of using a custom layout model versus a universal one is evaluated, further reducing segmentation errors. The study confirms that the HTR model can significantly enhance the processing of medieval Latin texts, offering efficient, accurate, and scalable transcription support for researchers in the Digital Humanities.