Kyiv-Mohyla
Humanities Journal
KYIV-MOHYLA SCHOLARLY PEER-REVIEWED JOURNALS

Handwritten Text Recognition of Ukrainian Manuscripts in the 21st Century: Possibilities, Challenges, and the Future of the First Generic AI-based Model

Author(s): Aleksej Tikhonov, Achim Rabus
Source: Kyiv-Mohyla Humanities Journal 11 (2024): 226–247
Published by: National University of Kyiv-Mohyla Academy

http://kmhj.ukma.edu.ua/

DOI: 10.18523/2313-4895.11.2024.226-247

Handwritten Text Recognition of Ukrainian Manuscripts in the 21st Century: Possibilities, Challenges, and the Future of the First Generic AI-based Model

Aleksej Tikhonov
University of Freiburg, University of Zurich

Achim Rabus
University of Freiburg

Abstract
This article reports on developing and evaluating a generic Handwritten Text Recognition (HTR) model created for the automatic computer-assisted transcription of Ukrainian handwriting publicly available via the HTR platform Transkribus. The model's training process encompasses diverse datasets, including historical manuscripts by renowned poets Taras Shevchenko and Lesya Ukrainka, along with private correspondence used for the General Regionally Annotated Corpus of Ukrainian (GRAC) and a diary procured at the Holodomor Museum collection. We evaluate the model's performance by comparing its theoretical accuracy, with a character error rate (CER) of 4.2%, against its practical efficacy when augmented with an AI-based language model for Ukrainian and a Large Language Model. The model is versatile and functional and can thus be applied for mass-digitization of Ukrainian cultural heritage. In our outlook section, we identify possibilities for further improving the model.

Key Words: Ukrainian, handwritten text recognition, manuscripts, handwriting, AI.

Introduction: Generic Handwritten Text Recognition Model for Ukrainian Manuscripts

Transcribing manuscripts is a challenging and time-consuming task not only due to the manual work itself but also due to variations in handwriting styles and individual differences caused by linguistic variation and change, spelling reforms, writing tools, the material condition of the manuscript, and many other factors. Existing Optical Character Recognition (OCR) models are tailored towards transcribing printed or typed text and cannot automatically recognize handwritten content. The manuscripts can be very different types of text: letters from well-known personalities or one's relatives or ancestors, novels or poems, diaries, or legal reports. There are almost no limits to the diversity of possible manuscripts as research objects but there are limits to human resources regarding large amounts of texts. Many in academia, archives, museums, initiatives, and other target groups will know what it means to work with hundreds or thousands of handwritten pages without corresponding digital transliterations. Such tasks require an

inordinate amount of patience, time, and skills in recognizing individual handwriting, often written using historical spelling or graphematic standards.

In the beginning, there is always the written word, contained in a manuscript, which has to be preserved for historical, political, or cultural reasons. To protect manuscripts, in most cases, a digital copy is used for research purposes; this is the basis for automatic Handwritten Text Recognition (HTR). In particular, manuscripts, the existence and exploration of which may be threatened by external factors such as natural disasters or war, require efficient processing and the broadest possible access for researchers, archivists, and other interested parties.

In 2022, the members of the MultiHTR project (Multilingual Handwritten Text Recognition)[1] at the University of Freiburg, Germany came up with the idea of creating a generic HTR model for modern Ukrainian. In the context of the project, modern Ukrainian refers to the time from the mid-19th century to the early 21st century. The motivation for developing such a model was to transcribe Ukrainian manuscripts from Ukraine and outside of Ukraine as efficiently as possible and thus make them accessible to a broader audience and to ensure further processing (e. g., translation into other languages, linguistic and historical studies, distant reading, etc.). The second motivation was to assist the Ukrainian-speaking diaspora in the Federal State of Baden-Württemberg and more broadly in Germany in preserving their cultural heritage by transcribing ego-documents (letters, recipes, lyrics, diaries, etc.). The model allows scanned or photographed handwritten texts to be automatically transliterated and thus makes them more accessible, regardless of the reading skills of an individual in specific handwriting styles.

The MultiHTR project published the first generic model for handwritten Ukrainian on the Transkribus platform[2] in May 2023.[3] The present article deals with the methods for model development, the compilation of the training data, the evaluation of the model and is, in fact, the first presentation of the model in an academic context.

Related Work: From an Artefact to a Digital Text

At present, numerous free and commercially licensed tools that provide OCR functionality for printed or typewritten texts are available. Prominent among these are Adobe Acrobat[4] or ABBYY FineReader.[5] Some applications (e. g.,

1 MultiHTR – Multilinguale Handschriftenerkennung. Projektbeschreibung. Accessed July 27, 2023. https://www.multihtr.uni-freiburg.de.

2 Unlock the past with Transkribus. Accessed July 27, 2023. https://www.transkribus.org.

3 Ukrainian generic handwriting: Free Public AI Model for Handwritten Text Recognition with Transkribus. Accessed July 27, 2023. https://readcoop.eu/model/ukrainian-generic-handwriting/.

4 Adobe Acrobat. Easily edit your scanned PDF documents with OCR. Accessed June 14, 2023. https://www.adobe.com/acrobat/how-to/ocr-software-convert-pdf-to-text.html.

5 Ukrainian language for ABBYY FineReader Professional Edition 8.0.1126.0. Accessed June 14, 2023. https://ukrainian-language-for-abbyy-finereader-professional-edition.updatestar.com/.

OCR4All[6]) also offer HTR approaches. Both types have a high accuracy in transmuting digital optical representations of text into editable formats commonly employed in word processors, such as .txt, .docx, and .odt. Furthermore, both types include support for recognizing Cyrillic characters as offered by OCR4All for handwritten texts since the early 2020s[7] and especially for printed or typewritten Ukrainian as exemplified by Adobe Acrobat and ABBYY FineReader since the late 2000s. Although the effective recognition of handwritten texts in any Cyrillic alphabet, particularly the Ukrainian one, has emerged as a salient area of interest in recent years, there is currently no universally valid methodology for such tasks, especially a methodology that is as barrier-free as possible from a technical point of view and does not require any programming knowledge.

The first Cyrillic handwritten text recognition experiment employed character-based Hidden Markov Models (HMMs). It yielded a recognition rate of 94.12%, specifically for limited keyword selections encompassing a small set of a few dozen items.[8] A similar attempt was made at recognizing words and lines in handwritten Cyrillic documents.[9] In 2008, Sergei Kornienko, Fedor Cherepanov and Leonid Iasnitckii proposed applying artificial intelligence approaches using neural network technologies to recognize the text of Old Church Slavonic manuscripts and early printed books. Overall, under ideal circumstances, they reached up to 80% correctly recognized glyphs for pre-modern Slavic manuscripts.[10] This was impressive for the time the project was conducted and a considerable step forward in comparison to the previous rather theoretical approaches, but the results have more theoretical than practical value, since, with an error rate of 20% of all characters at best, more or less every word would require manual correction. This, in turn, would slow the correction process to a level that renders the automatic pre-processing step impractical. Another approach focusing on the recognition of extended texts in the Cyrillic alphabet centered on Old Church Slavonic, which tackled diverse challenges in HTR, including the

6 Optical Character Recognition (and more) for everyone, accessed June 14, 2023, https://www.ocr4all.org/.

7 Alexander Winkler, "OCR4All Tools (Cyrillic)." HTML. 2020. Reprint, GitHub, 18 May 2021, https://github.com/alexander-winkler/ocr4all_tools.

8 M. D. Savic & M. Bojovic, "Recognition of Handwritten Text: Basic Concepts of a New Approach," in *4th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services, TELSIKS'99* (Cat. No.99EX365), 2:468–71, vol. 2, 1999, https://doi.org/10.1109/TELSKS.1999.806253.

9 Anoop M. Namboodiri & Anil K. Jain. "Online Handwritten Script Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, no. 1 (January 2004): 124–30. https://doi.org/10.1109/TPAMI.2004.1261096.

10 Sergei Kornienko, Fedor Cherepanov, & Leonid Iasnitckii, "Raspoznavanie tekstov rukopisnykh i staropechatnykh knig na osnove neirosetevykh tekhnologii" ["OCR of manuscripts and early printed books using neural networks"] (paper presented at conference "Modern Information Technologies and Written Heritage: from Ancient Texts to Electronic Libraries" – El'Manuscript-08, Kazan, Republic of Tatarstan, August 25–30, 2008). https://textualheritage.org/ru/el-manusctipt-08-/52.html.

erroneous identification of grapheme connections as one grapheme, the presence of decorative graphemes or elements, and all irregularities disrupting the uniform appearance of the old Cyrillic handwritten script.[11] More methodologies for recognizing old Slavic Cyrillic characters were introduced using decision trees and fuzzy classifiers based on shared character bitmap features and evaluating their efficiency and accuracy.[12] Achim Rabus[13] discussed the efficacy of neural network models in automatic text recognition of various styles of pre-modern Slavic handwriting on the Transkribus platform.[14] Notably, composite models amalgamating training data from diverse sources demonstrate the capability to transcribe disparate styles of old Cyrillic handwriting with error rates below 4%. Later, we released a model for handwritten Russian[15] within the MultiHTR project.

At the same time, during the late 2010s, a pioneering endeavor was undertaken to automate the transcription of handwritten Ukrainian. This undertaking was initiated by the Genealogical Society *RIDNI*, project, which presents a genealogical platform for Ukraine.[16] Within this project, a significant component involves digitizing and automatically recognizing records, such as birth and death certificates, and other predominantly tabular documents. Notably, this endeavor focuses on a restricted lexicon, primarily encompassing personal names, surnames, and places of birth or death. A different approach (Yevhen Bodnia and Mariia Kozulia, 2021, The National Technical University "Kharkiv Polytechnic Institute")[17] focuses on handwritten

11 Mimoza Klekovska, Igor Nedelkovski, Vera Stojcevska-Antic, & Dragan Mihajlov, "Automatic Letter Style Recognition of Churchslavic Manuscripts," in *Proceedings of Papers of the 44th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST 2009), Veliko Tarnovo, Bulgaria, June 25–27, 2009*, ed. by Rumen Arnaudov, vol. 1 (Sofia, 2009), 221–4.

12 Cveta Martinovska, Mimoza Klekovska, Igor Nedelkovski, & Dragan Kaevski, "Methodologies for Recognition of Old Slavic Cyrillic Characters," *International Journal of Computational Intelligence Studies* 2, no. 3–4 (January 2013): 264–87. https://doi.org/10.1504/IJCISTUDIES.2013.057639.

13 Achim Rabus, "Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus." *Scripta & E-Scripta*, 19 (2019): 9–32.

14 Church Slavonic (2): Free Public AI Model for Handwritten Text Recognition with Transkribus, accessed July 27, 2023, https://readcoop.eu/model/church-slavonic-2/.

15 Aleksej Tikhonov, Lesley Loew, Milanka Matić-Chalkitis, Martin Meindl, & Achim Rabus, "Multilingual Handwritten Text Recognition (MultiHTR) or Reading Your Grandma's Old Letters in German, Russian, Serbian, and Ottoman Turkish with Artificial Intelligence," in *The Palgrave Handbook of Digital and Public Humanities*, edited by Anne Schwan and Tara Thomson, 215–33. (Cham: Springer International Publishing, 2022). https://doi.org/10.1007/978-3-031-11886-9_12.

16 *Ridni:* henealohichne tovarystvo. Doslidzhennia rodovodu v Ukraini [*Ridni:* genealogical society. Research on genealogy in Ukraine], accessed: July 27, 2023, https://ridni.org/.

17 Bodnia, Yevhen, & Mariia Kozulia, "Web Application System of Handwritten Text Recognition," COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine. https://ceur-ws.org/Vol-2870/paper98.pdf.

Ukrainian data input into contemporary electronic devices such as smartphones and tablets. Their methodology deliberately bypasses the intermediate stage encompassing the materiality of the manuscripts and, thus, exhibits a forward-looking perspective. However, regarding material and more extensive manuscripts, the Transkribus platform (and similar platforms such as eScriptorium[18] presents a valuable technology for manuscripts housed within archival repositories, museums, and libraries in different languages. The 'freemium' platform Transkribus has been available for some time and is a web- and server-based GUI tool for transcribing manuscripts. Its most interesting feature is the possibility to train models for automatic (pre-)transcription of handwritten text in different languages and scripts. The Handwritten Text Recognition (HTR) technology implemented in Transkribus is based on artificial intelligence and neural networks. According to the Transkribus FAQ page, "Unlike OCR, HTR does not focus on individual letters. Instead, it scans and processes the image of entire lines and tries to decode this data."[19] Across the globe, repositories hold substantial collections of handwritten archival materials, eagerly anticipating the day when they will undergo digitization and become accessible to the public, or at the very least, to the academic community. Regrettably, manuscripts that are yet to be digitized are in danger of being destroyed by natural disasters or wars, causing irreparable loss. This holds especially true for Ukrainian manuscripts, which face potential damage due to the ongoing Russian war of aggression against Ukraine. Given this tragic reality, the MultiHTR project addressed the need to train a model in early 2022, specifically for Ukrainian handwritten texts. After commencing active work in the summer of 2022, we released the first generic Transkribus model for handwritten Ukrainian that can be used to transcribe manuscripts from the late 19th to the early 21st centuries in May 2023.

The First Generic Ukrainian Handwriting Model

In the subsequent sections, we explain the process of training a model for Handwritten Text Recognition (HTR). Furthermore, we provide a comprehensive account of the specific data employed for training a generic model for Ukrainian. The training procedure encompasses a series of steps aimed at optimizing the model's performance in accurately transcribing handwritten texts. By elucidating the methodology employed and detailing the precise data sources harnessed in the training process, we seek to offer a comprehensive understanding of the development and training of the HTR model for handwritten Ukrainian.

---

18 Scripta / escriptorium. GitLab, accessed July 27, 2023, https://gitlab.com/scripta/escriptorium.

19 What is the difference between OCR (Optical Character Recognition) and HTR? https://readcoop.eu/transkribus/help/what-is-the-difference-between-ocr-optical-character- recognition-and-htr/.

Methodology

Training a neural network model serves as an illustrative example of supervised machine learning, wherein the acquisition of a specific quantity of training data becomes imperative for successful model training. In the context of the HTR-training, this means that digital representations of the manuscript to be transcribed need to be available, together with an accompanying textual transcription that aligns precisely (i.e., line by line) with the original manuscript text. Within disciplines such as history and literature, unintentional errors or grammatical inaccuracies admitted by the scribe found in the original manuscript during the transcription process are usually corrected. However, a precise one-to-one correspondence between the written language data and the original source material leads to the best results. In instances where the trainer encounters passages within the text that exhibit genuine obscurity or ambiguity, it is admissible to annotate such sections as 'unclear,' thereby prompting the algorithm to disregard them during the training. Similarly, crossed-out regions can be marked as such and are duly incorporated into the training procedure as crossed-out written data.

Additionally, it is possible to train so-called "smart" models. Here, the model learns to resolve frequent abbreviations and address systematic graphematic and grammatical errors evident in handwritten text. However, it is essential to note that the subject matter of this present article centers on a specific one-to-one transliteration model, rather than encompassing the broader spectrum of smart models.

The efficacy of the model is contingent upon the volume of training data, as a larger corpus typically leads to enhanced model accuracy. As per the guidelines provided in the Transkribus FAQs, a minimum of 15,000 transcribed words is recommended to develop a functional model. Notably, it is possible to train models for numerous languages and scripts. Through the processing of characters, words, and lines, and comparison between transcriptions and corresponding digital images, the model acquires the ability to accurately transcribe handwritten graphemes. This requires multiple iterations, known as epochs,[20] involving the model's statistical prediction evaluation against the correct data. The more exposure the model has to the transcribed data, the more adeptly it adapts to distinctive handwriting styles. In the Transkribus platform, the maximum value for training models using the integrated PyLaia engine is 250 epochs.

Figure 1 provides a representative depiction of the learning curve for one of the training steps of the handwritten Ukrainian Transkribus model based on the PyLaia technology, highlighting a typical progression:

20 During training, the model is presented with smaller subsets of data at a time, and it performs forward propagation to make predictions. Then, the model calculates the errors between its predictions and the ground-truth. The learning rate plays a crucial role in shaping the training progression, determining the increment from one epoch to the next. A higher learning rate results in a swifter decrease in the CER.

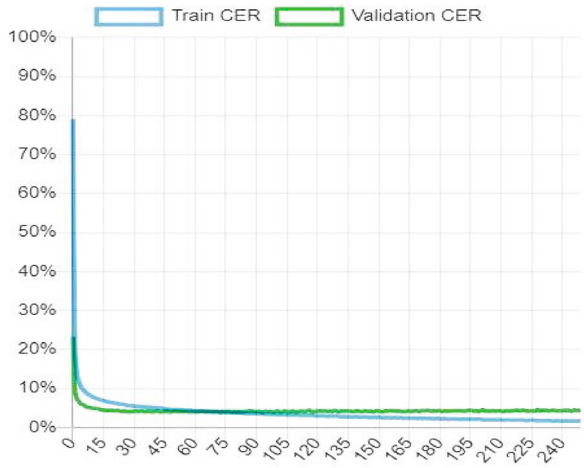

Figure 1: Progression of a learning curve of the HTR(PyLaia)-algorithm

The computed accuracy is indicated by the so-called Character Error Rate (CER). The model depicted in Figure 1, trained on a combination of numerous different Ukrainian hands (the exact description of the text used will be provided in the next section), has a CER of 1.71% on the training set and 4.5% on the test set after 250 training epochs. In machine learning, the performance of a model is usually evaluated on a test or validation set that is part of the overall data but has not been seen during training. In the specific case of training handwritten text recognition models with Transkribus, certain pages of the manuscript(s) used for model training are set aside as a test or validation set. Naturally, the performance using the training set (i. e., data seen during training) is usually better than the performance using the test or validation set (i.e., data not seen during training). As one can see, during the initial 10 or so epochs, CER drops drastically with both training and test data. After that, CER drops less and less significantly with each additional epoch. Therefore, a typical neural network model training curve has a hyperbolic shape.

Pre-processing and uploading

For optimal transcription, it is essential to have digital copies of manuscripts in the form of high-quality color scans or photographs available. When employing photographs, it is critical to ensure an approximate angle of 90° to the manuscript to prevent distortion of aspect ratios and maintain the integrity of the manuscript image. Such digital copies may be uploaded to the Transkribus platform for training (or transcription) in either *.jpeg or *.pdf format. Based on experience, files containing up to 1,000 pages typically present no issues during uploading and processing. Nonetheless, at the outset, it is advisable to upload a small set of sample data, comprising 1 to 25 pages, to facilitate the selection of the appropriate layout analysis (LA, see below) and transcription model settings.

Crucially, all computational processes occur on the Transkribus servers in Austria, freeing up users' personal computers or servers, which remain unburdened

and available for other tasks. As a result, the computer doesn't need to remain operational and online during lengthier automated processes, such as training new transcription models., which may last several hours or even days.

Dataset and Training

There are two prototypical categories of HTR-models: (a) specific models and (b) generic models. For specific models, the training sets are composed of homogeneous data, primarily employed to transcribe extensive collections of manuscripts with similar handwriting. These collections typically encompass personal notes or manuscripts exhibiting slight disparities between a limited number of scribes, resembling printed text such as Middle High German or other medieval manuscripts. In contrast, for generic models, the objective is to forge a versatile model capable of deciphering diverse paleographic or individual handwriting styles. Consequently, the data sets must be heterogeneous, encompassing a wide array of handwriting styles. However, it is imperative to maintain a degree of selectiveness and focus on a specific historical period. As a rule, one needs more training data to successfully train a generic model.

In the context of this article, the selected period spans from the latter half of the 19th century to the late 20th century. The overarching goal is to train a generic model for Ukrainian, capable of automatically transcribing relatively modern texts. To attain this goal, we compiled a training data set comprising manuscripts from the 19th up to the 21st century.

Figure 2: The first trial attempt of training with Ukrainian handwritten data «Shev_Kulish 1"

Creating a generic Ukrainian model entailed searching for suitable data, as the project commenced without any existing manuscript resources. We started our

investigation with an exploration of the SUCHO platform (Saving Ukrainian Cultural Heritage Online),[21] an international volunteer initiative conceived by Anna Kijas (Tufts University, US), Quinn Dombrowski (Stanford University, US), and Sebastian Majstorovic (European University Institute, Italy). This platform proved instrumental thanks to its extensive compilation of over 5,000 websites and databases, facilitating the discovery of the first appropriate manuscripts.

The inaugural manuscript chosen for model training was avtohraf of Taras Shevchenko's poem *Posadzhu kolo khatyny…* (*I will plant near the hut…*).[22] The selection was motivated by the manuscript's accessibility; it boasts a publicly available transcription either in its entirety or in part. The significance of Shevchenko's works as a cornerstone of Ukrainian national poetry was another reason why it was deemed an ideal candidate. The process of preparing the text for training involved aligning the manuscript images with the corresponding digital transcriptions found online, thereby initiating the development of our first specific model named "Shev_Kulish 1":

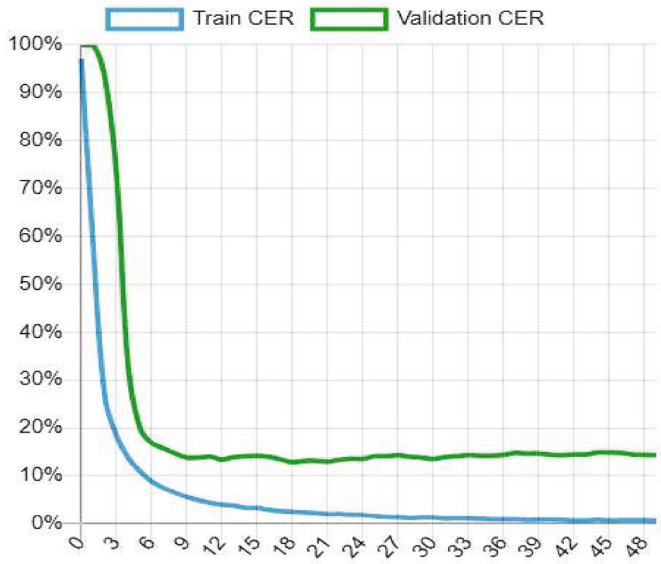

Figure 3: The second attempt of training with Ukrainian handwritten data "Shev_Kulish 2"

The first model, exhibiting a CER of 10.79%, was successfully trained utilizing 861 word forms and 191 lines. Consequently, it can be inferred that approximately 89.21% of the automatic transliterations would accurately correspond to the original text if other manuscripts by Shevchenko/Kulish were subjected to transliteration.

21 Saving Ukrainian Cultural Heritage Online, accessed August 3, 2023, https://www.sucho.org/.

22 Taras Shevchenko, "Posadzhu kolo khatyny…": virsh, rukopys, bilovyi avtohraf ostannoi redaktsii virsha "Podrazhaniie" v rukopysnomu zbirnyku P. O. Kulisha ["I Will Plant Near the Hut…": poem, manuscript, clean autograph of the last edition of the poem "Imitation" in the manuscript collection of P. O. Kulish], 187?-188?, no. 28438, fond I: Literaturni materialy [Literary materials]. Manuscript Institute of the V. I. Vernadskyi National Library, Kyiv.

Subsequently, the model was employed to transliterate some segments of the manuscript for which no transcription was available. Here, approximately 10.79% of erroneously transliterated passages were manually corrected. The resulting dataset, comprising the complete manuscript alongside its corresponding transliterations, served as the foundation for developing the second specific model, named "Shev_Kulish 2" (trained with 2,266 word forms):

The higher Character Error Rate (CER) of 14.3% of the second model can be attributed to the greater volume of data incorporated, leading to an increased likelihood of encountering diverse spellings of graphemes, letter forms, and grapheme combinations. Consequently, CER tends to rise after the first model, and with the inclusion of each subsequent new text. To this end, another handwritten text, Taras Shevchenko *Topolia ta inshi virshi: Zbirnyk poezii*[23]*, was transliterated using "Shew_Kulish 2." Following manual error correction, both manuscripts were employed in the training of a new model: "Ukrainian_Shev_Kulish_Topol_1" (trained with 4,960-word forms):*

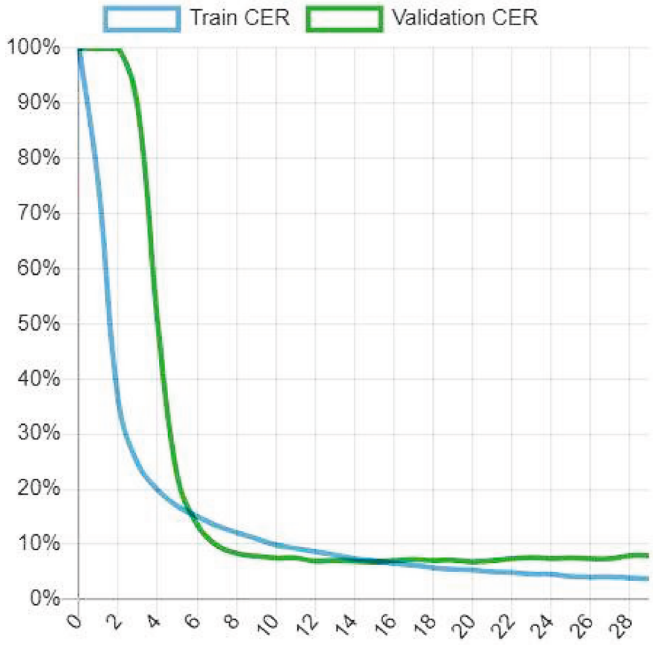

Figure 4: The learning curve of the first model which includes more than one Ukrainian handwritten text "Ukrainian_Shev_Kulish_Topol_1"

Subsequently, as the CER for older training texts improved to 7.95%, we embarked on a parallel trajectory, initiating the compilation of a second diachronically more recent dataset sourced from the General Regionally Annotated Corpus of Ukrainian Language (GRAC)[24]. This dataset, curated by Maria Shvedova, Ruprecht von Waldenfels, Serhij Yaryhin, Andriy Rysin, Vasyl Starko, Tymofij Nikolajenko, et al. (2017–2023), was

23 Taras Shevchenko, "Topolia" ta inshi virshi, zbirnyk poezii, rukopysnyi spysok ["Topolia" and other poems, a collection of poems, handwritten list], 18??, no. 7448, fond I: Literaturni materialy [Literary materials]. Manuscript Institute of the V. I. Vernadskyi National Library, Kyiv.

24 GRAC, accessed August 3, 2023. https://uacorpus.org/Kyiv/ua.

used for further training. Encompassing a compilation of letters and postcards written by at least 30 distinct individuals in the 1980s-early 2000s and transcribed by students of Lviv Polytechnic University between 2018 and 2020, the dataset augments the model's capacity to encompass contemporary linguistic and handwriting- specific variation. These manuscripts comprise only a small segment within the GRAC corpus, which contains various kinds of texts.

By combining the two Shevchenko texts, the GRAC letters, and an additional diary from the 1960s written by Lavrin Nechyporenko and provided by The National Museum of the Holodomor Genocide in Kyiv, we created the first non-public generic model, "Generic_Ukrainian_0.01" (trained with 71,583 word forms):

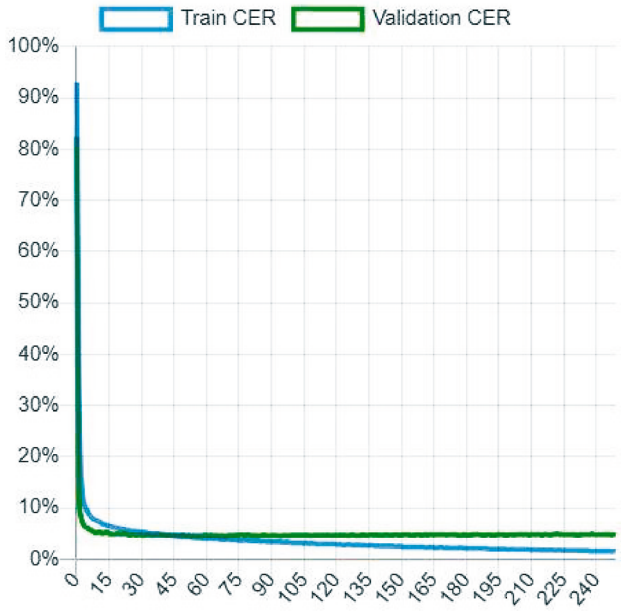

Figure 5: The learning curve of the first attempt for a generic Ukrainian HTR-moel "Generic_Ukrainian_0.01"

"Generic_Ukrainian_0.01" is a promising generic model with a CER of 4.7%, attesting to its theoretical accuracy. Nevertheless, prevailing research findings show that a truly functional generic HTR model typically needs to be trained on a substantial corpus, ranging from 100,000 to 250,000 word forms[25]. Acknowledging this requirement, we further expanded the model by incorporating one more manuscript, lithographed

25 Guenter Muehlberger, et al., "Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study," *Journal of Documentation* 75, no. 5 (2019): 954–76; Burlacu, Constanța, and Achim Rabus. "Digitising (Romanian) Cyrillic Using Transkribus: New Perspectives." *Diacronia* 14 (December 12, 2021): A196(1–9). Tikhonov, Aleksej, Lesley Loew, Milanka Matić-Chalkitis, Martin Meindl, & Achim Rabus. "Multilingual Handwritten Text Recognition (MultiHTR) or Reading Your Grandma's Old Letters in German, Russian, Serbian, and Ottoman Turkish with Artificial Intelligence." In *The Palgrave Handbook of Digital and Public Humanities*, edited by Anne Schwan and Tara Thomson, 215–33 (Cham: Springer International Publishing, 2022).

collection of folk melodies from the voice of Lesya Ukrainka.[26] Incorporating the additional text, the model achieved a comprehensive and representative scope, leading to its being officially published in May 2022 with the title "Ukrainian generic handwriting 1"15 (trained with 100,079 word forms)[27]:

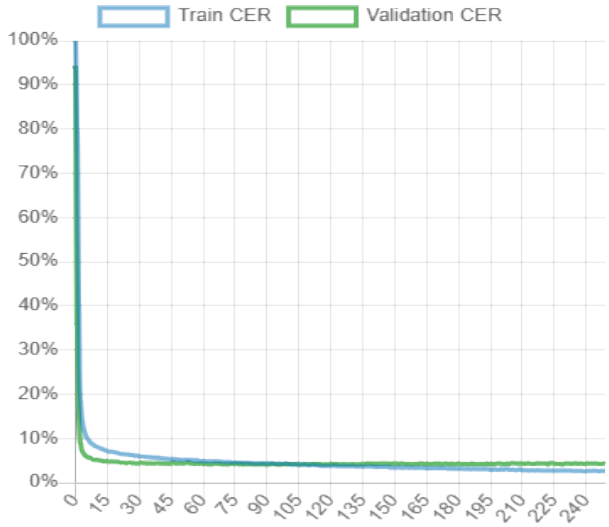

Figure 6: The learning curve of the first publicly available generic HTR-model for Ukrainian "Ukrainian generic handwriting 1"

With a CER of 4.2%, the first public Ukrainian HTR model provides a robust and versatile method, capable of accommodating a wide range of handwriting styles across the Ukrainian language in the 19th and 21st centuries. It can be used via the Transkribus platform free of charge by anyone interested. However, the CER value pertains to the theoretical performance of the model. To gauge its practical efficacy, we report on some tests to evaluate how the model handles actual handwritten texts that lie outside the purview of the training set and originate from scribes unknown to the model. This evaluation will shed light on the model's adaptability and generalizability to novel handwriting styles and linguistic variation, thereby assessing its real-world functionality and overall performance.

Evaluation

For the evaluation of the public model, a Telegram post authored by Ukrainian President Volodymyr Zelenskyi on July 14, 2022,[28] was copied by hand using black and

26 Klyment Kvitka, ed., *Narodni melodii. Z holosu Lesi Ukrainky* [*Folk Melodies. From the* Voice of Lesya Ukrainka], vol. 1 (Kyiv, 1917).

27 See: https://readcoop.eu/model/ukrainian-generic-handwriting/.

28 Volodymyr Zelenskyi, "Vziav uchast u konferentsii v Haazi, meta yakoi – poriatunok mizhnarodnoho prava…" ["Participated in a conference in The Hague, the goal of which was to save international law…"], July 14, 2022, https://t.me/V_Zelenskiy_official/2534, accessed: July 25, 2023.

blue ink on white paper and subsequently scanned. Notably, we deliberately overlooked the linearity of line progressions and the uniformity of graphemes during the copying process, aiming to present the model with a realistic example for processing.

Before automatic HTR, Layout Analysis (LA) has to be conducted within Transkribus. It entails identifying text and line regions, as well as baselines. While this step generally proceeds smoothly, certain manuscript types may present complexities. For instance, unusually expansive inter-word distances in manuscripts can inadvertently lead to individual words being identified as distinct lines, thereby impeding automated transliteration. However, in scenarios akin to the training procedure for the Ukrainian model and the test cases discussed in this article, we encounter normal word spacing and standardized layout attributes, such as predominantly horizontal lines. In our test case, Layout Analysis unfolded seamlessly, with all lines being recognized accurately (*Figure 7*):

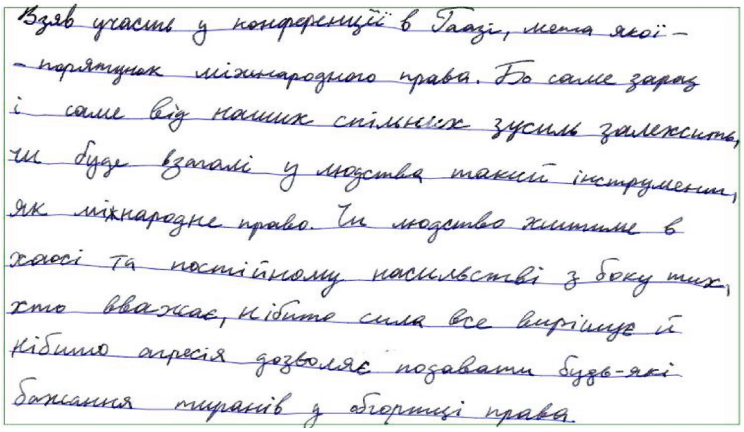

Figure 7: Evaluation manuscript 1

Regarding automatic HTR, the model performed rather well. However, several errors could be observed (*Figure 8*):

1-1 # Взяв участь ~~з~~ у конференції в ~~Газі, мена~~ Гаазі, мета якої –
1-2 # - порятунок міжнародного права. ~~До~~ Бо саме зараз
1-3 # і ~~сами~~ саме від наших спільних зусиль залежить,
1-4 # чи буде взагалі у людства такий ~~інструменний~~ інструмент,
1-5 # ~~Як міхнародне~~ як міжнародне право. Чи людство ~~житили~~ житиме в
1-6 # ~~хобі~~ хаосі та постійному насильстві з боку ~~тих~~ тих,
1-7 # хто вважає, нібито сила все вирішує й
1-8 # ~~Нібито~~ нібито агресія дозволяє подавати будь-які
1-9 # бажання тиранів ~~з~~ у обгортці права.

Figure 8: Comparison between the HTR transcription of evaluation manuscript 1 (Figure 7) and the original wording

Words marked in red signify inaccuracies that occurred during the automatic HTR process. Conversely, the green areas show the correct reading. Upon initial inspection, it may seem that our public "Ukrainian generic handwriting 1" model

produced numerous errors. However, upon closer examination, it becomes evident that the majority of the manual corrections pertain to individual signs (graphemes, punctuation, numbers). As part of the automatic comparison process in Transkribus, the complete word form is marked as incorrect, even in the case of merely one incorrect sign. Thus, if the exact CER for this specific practical example is computed, the following result is obtained (*Table 1*):

Line	N signs	N incorrect signs	individual absolute CER
1	38	3 (4)	7.89 %
2	38	1	2.63 %
3	36	1	2.77 %
4	37	3	8.11 %
5	34	3	8.82 %
6	36	4	11.11 %
7	31	0	0 %
8	37	0 (1)	0 %
9	29	1	3.45 %
Total	331	16	4.83 %

Table 1: Evaluation of the real-life performance "Ukrainian generic handwriting 1"

Despite some errors, this real-world example shows that the practical CER is almost as low as the computed CER.

In Transkribus, users have the option of transliteration without a language model (as in the experiment just reported), or with a language model generated from the training data. Enabling the language model means that, during transliteration, linguistic features (e.g., words) will be taken into account to a greater extent. Enabling the language model for the transliteration of the manuscript shown in Figure 7 yields the following result (Figure 9):

1-1 # Взяв участь ~~з~~ у конференції в ~~Газі, мена~~ Гаазі, мета якої ~~—~~
1-2 # - порятунок ~~міжнародною~~ міжнародного права. ~~До~~ Бо саме зараз
1-3 # і ~~сами~~ саме від наших спільних зусиль залежить,
1-4 # чи буде взагалі у людства такий ~~інструменний~~ інструмент,
1-5 # як ~~міхнародне~~ міжнародне право. Чи людство ~~житили~~ житиме в
1-6 # ~~холосі~~ хаосі та постійному насильстві з боку ~~тих~~ тих,
1-7 # хто вважає, нібито сила все вирішує й
1-8 # ~~Нібито~~ нібито агресія дозволяє подавати будь-які
1-9 # бажання тиранів ~~з~~ у обгортці права.

Figure 9: Comparison between the HTR transcription (with the language model from the training data) of evaluation manuscript 1 (Figure 7) and the original wording

As can be seen, in this instance, the results are – somewhat unexpectedly – slightly, albeit insignificantly worse than when using the model without a language model.

In an additional experiment, we used OpenAI's GPT-4 (chat.openai.com), an improved version of the famous generative AI chatbot ChatGPT and arguably the most advanced general-purpose Large Language Model at the time of writing this article, for post-correction of the HTR transcription. We prompted GPT-4 to make gentle corrections to the HTR errors in the text and obtained the following results:

1-1 # Взяв участь у конференції в ~~Газі, назва~~ Гаазі, мета якої ~~—~~ —
1-2 # ~~-~~ порятунок міжнародного права. ~~То~~ Бо саме зараз
1-3 # і ~~самі~~ саме від наших спільних зусиль залежить,
1-4 # чи буде взагалі у людства такий ~~інструмент.~~ інструмент,
1-5 # ~~Як~~ як міжнародне право. Чи людство житиме в
1-6 # хаосі та постійному насильстві з боку тих,
1-7 # хто вважає, нібито сила все вирішує й
1-8 # нібито агресія дозволяє ~~задовольнити~~ подавати будь-які
1-9 # бажання тиранів ~~під обгорткою~~ у обгортці права.

Figure 10: Automatic correction of the HTR transcription (with the language model from the training data) of evaluation manuscript 1 (Figure 7) with GPT-4

As can be seen, GPT-4 was able to correct some obvious errors such as *інструмент* 'instrument' (with incorrect punctuation) and *міжнародне* 'international'. However, it added some unnecessary errors such as *задовольнити* 'satisfy', instead of *подавати* 'grant' and *під обгорткою* instead of *у обгортці* 'in the guise of'.

Another version of the GPT4-correction with a slightly different prompt yielded the following result:

1-1 # Взяв участь у конференції в ~~Газі, назва~~ Гаазі, мета якої ~~—~~ —
1-2 # ~~"Порятунок -~~ порятунок міжнародного ~~права". То~~ права. Бо саме зараз
1-3 # і ~~самі~~ саме від наших спільних зусиль залежить,
1-4 # чи буде взагалі у людства такий інструмент,
1-5 # як міжнародне право. Чи людство житиме в
1-6 # хаосі та постійному насильстві з боку тих,
1-7 # хто вважає, нібито сила все ~~вирішує,~~ вирішує й
1-8 # ~~й~~ нібито агресія дозволяє ~~задавати~~ подавати будь-які
1-9 # бажання тиранів ~~з обгорткою~~ у обгортці права.

Figure 11: Automatic correction of the HTR transcription (with the language model from the training data) of the evaluation manuscript 1 (Figure 7) with GPT-4

Here, punctuation and capitalization in lines 1-4 and 1-5 are correct. While still erroneous, *задавати* 'to ask' instead of *подавати* 'to provide' (1-8) and з обгорткою 'with a wrapper' instead of під обгорткою 'under a wrap' (1-9) are arguably better. Interestingly, in both instances, the model insisted that *обгортці* 'a wrap' in the locative case instead of the instrumental case is not possible here.

In another experiment, we used different handwriting with the same text (*Figure 12*) and had it transcribed by our public Transkribus model (without the language model enabled):

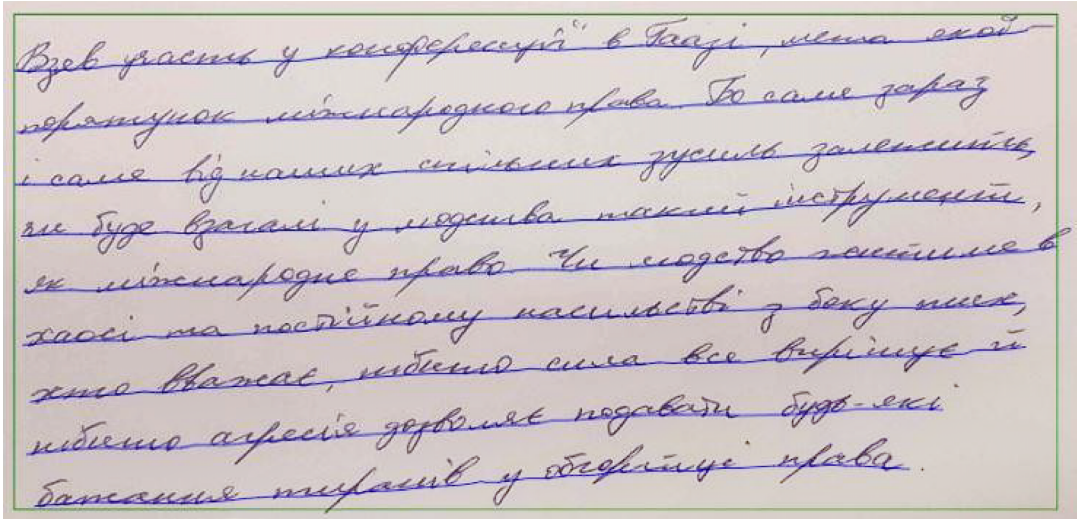

Figure 12: Evaluation manuscript 2

As can be seen, the results are slightly worse than with the first hand:

1-1 # ~~Взев~~ Взяв участь у ~~когференції~~ конференції в ~~Раазі,~~ Гаазі, мета ~~якод~~ якої –
1-2 # ~~-~~ порятунок ~~міжнародною~~ міжнародного права. Бо ~~саля~~ саме зараз
1-3 # і саме від наших спільних зусиль ~~заленить,~~ залежить,
1-4 # чи буде ~~взагам~~ взагалі у людства такий ~~йструмент~~ інструмент,
1-5 # як міжнародне право. Чи людство житиме в
1-6 # хаосі та постійному насильстві з ~~баку~~ боку тих,
1-7 # ~~кто~~ хто вважає, ~~піднто~~ нібито сила все ~~вирінує~~ вирішує й
1-8 # нібито агресія дозволяє подавати будь-які
1-9 # ~~Бажання тихашів~~ бажання тиранів у ~~очертце права~~ обгортці права.

Figure 13: Comparison between the HTR transcription of evaluation manuscript 2 (Figure 12) and the original wording

Quantitatively, the results look as follows:

Line	N signs	N incorrect signs	individual absolute CER
1	38	4	10.5 %
2	38	4 (5)	10.5 %
3	36	1	2.77 %
4	37	5	13.5 %
5	34	0	0 %
6	36	1	2.77 %
7	31	4	12.9 %
8	37	0	0 %
9	29	7 (8)	24.14 %
Total	331	26	7,90 %

Table 3: Evaluation of the real-life performance "Ukrainian generic handwriting 1" on the same text (Footnote 16), but a different scribe

Using our HTR model with the implemented language enabled yields the following results:

1-1 # Взяв участь у ~~когерекції~~ конференції в ~~Разі,~~ Гаазі, мета ~~якод =~~ якої –
1-2 # ~~-~~ порятунок ~~міжнародною~~ міжнародного права. Бо ~~саля~~ саме зараз
1-3 # і саме від наших спільних зусиль залежить,
1-4 # чи буде ~~взагам~~ взагалі у людства такий ~~струмент~~ інструмент,
1-5 # як міжнародне право. Чи людство житиме в
1-6 # хаосі та постійному насильстві з боку тих,
1-7 # хто вважає, ~~піднто~~ нібито сила все вирішує й
1-8 # нібито агресія дозволяє подавати будь-які
1-9 # бажання ~~тихалів~~ тиранів у ~~очерце права~~ обгортці права.

Figure 14: Comparison between the HTR transcription of evaluation manuscript 2 (Figure 12) with the language model and the original wording

As opposed to HTR without the language model enabled, some obvious errors such as *Взев* instead of *взяв* 'by participating' (1-1), *заленить* instead of *залежить* 'depends on', *баку* instead of *боку* 'side' (1-6) or кто instead of хто 'who' (1-7) have been corrected, while other errors remain. Quantitatively, the results are as follows:

Line	N signs	N incorrect signs	individual absolute CER
1	38	6	15.78 %
2	38	4 (5)	10.5 %
3	36	0	0 %
4	37	4	10.81 %
5	34	0	0 %
6	36	0	0 %
7	31	2	6.45 %
8	37	0	0 %
9	29	7	24.14 %
Total	331	23	6.94 %

Table 4: Evaluation of the real-life performance of "Ukrainian generic handwriting 1" (incl. language model) on the same text (Footnote 16) but with a different scribe

As can be seen, the improvement over the version without the language model is higher with this handwriting than with the handwriting discussed above. Apparently, the language model is more effective when the underlying HTR model struggles with the handwriting.

Prompting the Large Language Model GPT-4 to correct this text yields the following results:

1-1 # Взяв участь у конференції в ~~Римі,~~ Гаазі, мета якої –
1-2 # ~~-~~ порятунок міжнародного права. Бо саме зараз
1-3 # і саме від наших спільних зусиль залежить,
1-4 # чи буде взагалі у людства такий ~~інструмент~~ інструмент,
1-5 # як міжнародне право. Чи людство житиме в
1-6 # хаосі та постійному насильстві з боку тих,
1-7 # хто вважає, ~~що~~ нібито сила все вирішує й
1-8 # нібито агресія дозволяє подавати будь-які
1-9 # бажання ~~стихії в очереті~~ тиранів у обгортці права.

Figure 15: Comparison between the HTR transcription of the evaluation manuscript 2 (Figure 12), incl. the language model + GPT4 corrections, and the original wording

The results are considerably better with the correct rendition of *конференції* 'conference', *міжнародного* 'international', *взагалі* 'in general', *інструмент* 'instrument' (apart from the comma). However, line 1-9 is almost completely wrong and *що* 'what' in line 1-7 changes the connotation of the sentence. Overall, while this issue needs some further research, Large Language Models such as GPT4 are a promising way to improve less-than-ideal HTR results in a separate post-processing step.

Outlook. The Second Ukrainian Handwriting Model

Despite the detected disparities, it could be shown that the model is capable of transcribing handwritten Ukrainian texts with a low error rate. The proportion of correctly transliterated text is consistently above 92%. While the quality might vary depending on which handwriting to transcribe, the model's overall effectiveness remains high, offering a reliable and capable solution for Handwritten Text Recognition tasks and, thus, for mass digitization. We encourage all scholars concerned with the Ukrainian handwritten cultural heritage to test our model and to contact us, should any questions arise.

In cases where the Transkribus model underperforms, it might be a worthwhile undertaking to experiment with Large Language Models such as GPT-4 for the post-correction of mediocre HTR results.

The release of the Transkribus model marks a significant milestone for working with modern Ukrainian manuscripts. We plan to publish a second, improved version of our model of Ukrainian handwriting in the near future. To make this model more versatile, a wide variety of sources with different handwriting styles that can be used for training purposes is essential. Because of that, we appeal to our fellow researchers – both inside and outside of Ukraine – who have access to handwritten text and corresponding transcriptions that can be used for training purposes to get in touch.

We are dedicated to further enhancing the model's capabilities, aiming to deliver an even more comprehensive solution for HTRizing Ukrainian manuscripts from the 19th–21st centuries and contributing to the ever-evolving landscape of language technology for Ukrainian studies.

Bibliography

Adobe Acrobat. Easily edit your scanned PDF documents with OCR. Accessed June 14, 2023. https://www.adobe.com/acrobat/how-to/ocr-software-convert-pdf-to-text.html.

Bodnia, Yevhen, & Mariia Kozulia. "Web Application System of Handwritten Text Recognition." *COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine*. https://ceur-ws.org/Vol-2870/paper98.pdf.

Burlacu, Constanța, & Achim Rabus. "Digitising (Romanian) Cyrillic Using Transkribus: New Perspectives." *Diacronia* 14 (December 12, 2021): A196(1–9). https://doi.org/10.17684/i14A196en.

Church Slavonic (2): Free Public AI Model for Handwritten Text Recognition with Transkribus, accessed July 27, 2023, https://readcoop.eu/model/church-slavonic-2/.

GRAC, accessed August 3, 2023. https://uacorpus.org/Kyiv/ua.

Klekovska, Mimoza, Igor Nedelkovski, Vera Stojcevska-Antic, & Dragan Mihajlov. "Automatic Letter Style Recognition of Churchslavic Manuscripts." In *Proceedings of Papers of the 44th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST 2009), Veliko Tarnovo, Bulgaria, June 25–27, 2009*. Vol. 1, 221–4 (Sofia, 2009).

Kornienko, Sergei, Fedor Cherepanov, & Leonid Iasnitckii. "Raspoznavanie tekstov rukopisnykh i staropechatnykh knig na osnove neirosetevykh tekhnologii" ["OCR of manuscripts and early printed books using neural networks"]. Paper presented at conference "Modern Information Technologies and Written Heritage: from Ancient Texts to Electronic Libraries" – El'Manuscript-08, Kazan, Republic of Tatarstan, August 25–30, 2008. https://textualheritage.org/ru/el-manusctipt-08-/52.html.

Klyment, Kvitka, ed. *Narodni melodii. Z holosu Lesi Ukrainky* [*Folk Melodies. From the* Voice of Lesya Ukrainka]. Vol. 1. Kyiv, 1917.

Martinovska, Cveta, Mimoza Klekovska, Igor Nedelkovski, & Dragan Kaevski. "Methodologies for Recognition of Old Slavic Cyrillic Characters." *International Journal of Computational Intelligence Studies* 2, no. 3–4 (January 2013): 264–87. https://doi.org/10.1504/IJCISTUDIES.2013.057639.

Muehlberger, Guenter et al. "Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study." *Journal of Documentation* 75, no. 5 (2019): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

MultiHTR – Multilinguale Handschriftenerkennung. Projektbeschreibung. Accessed July 27, 2023. https://www.multihtr.uni-freiburg.de.

Namboodiri, Anoop M., & Anil K. Jain. "Online Handwritten Script Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, no. 1 (January 2004): 124–30. https://doi.org/10.1109/TPAMI.2004.1261096.

Optical Character Recognition (and more) for everyone, accessed June 14, 2023, https://www.ocr4all.org/.

Rabus, Achim. "Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus'. *Scripta & E-Scripta* 19 (2019): 9–32.

Ridni: henealohichne tovarystvo. Doslidzhennia rodovodu v Ukraini [*Ridni:* genealogical society. Research on genealogy in Ukraine]. Accessed: July 27, 2023, https://ridni.org/.

Savic, M. D., & M. Bojovic. "Recognition of Handwritten Text: Basic Concepts of a New Approach." In *4th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services.* TELSIKS'99 (Cat. No.99EX365), 2:468–71, vol. 2, 1999. https://doi.org/10.1109/TELSKS.1999.806253.

Saving Ukrainian Cultural Heritage Online. Accessed August 3, 2023, https://www.sucho.org/.

Scripta / escriptorium. GitLab, accessed July 27, 2023. https://gitlab.com/scripta/escriptorium.

Shevchenko, Taras. "Posadzhu kolo khatyny...": virsh, rukopys, bilovyi avtohraf ostannoi redaktsii virsha "Podrazhaniie" v rukopysnomu zbirnyku P. O. Kulisha ["I Will Plant Near the Hut...": poem, manuscript, clean autograph of the last edition of the poem "Imitation" in the manuscript collection of P. O. Kulish]. 187?-188? No. 28438. Fond I: Literaturni materialy [Literary materials]. Manuscript Institute of the V. I. Vernadskyi National Library, Kyiv. http://irbis-nbuv.gov.ua/dlib/item/0000613.

— — —. "Topolia" ta inshi virshi. Zbirnyk poezii. Rukopysnyi spysok ["Topolia" and other poems. A collection of poems. Handwritten list]. 18?? No. 7448. Fond I: Literaturni materialy [Literary materials]. Manuscript Institute of the V. I. Vernadskyi National Library, Kyiv.

Tikhonov, Aleksej, Lesley Loew, Milanka Matić-Chalkitis, Martin Meindl, & Achim Rabus. "Multilingual Handwritten Text Recognition (MultiHTR) or Reading Your Grandma's Old Letters in German, Russian, Serbian, and Ottoman Turkish with Artificial Intelligence." In *The Palgrave Handbook of Digital and Public Humanities*, edited by Anne Schwan and Tara Thomson, 215–33. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-11886-9_12.

Ukrainian generic handwriting: Free Public AI Model for Handwritten Text Recognition with Transkribus. Accessed July 27, 2023. https://readcoop.eu/model/ukrainian-generic-handwriting/.

Ukrainian language for ABBYY FineReader Professional Edition 8.0.1126.0. Accessed June 14, 2023. https://ukrainian-language-for-abbyy-finereader-professional-edition.updatestar.com/.

Unlock the past with Transkribus. Accessed July 27, 2023. https://www.transkribus.org.

What is the difference between OCR (Optical Character Recognition) and HTR? https://readcoop.eu/transkribus/help/what-is-the-difference-between-ocr-optical-character-recognition-and-htr/.

Winkler, Alexander. "OCR4All Tools (Cyrillic)." HTML. 2020. Reprint, GitHub, 18 May 2021. https://github.com/alexander-winkler/ocr4all_tools.

Zelenskyi, Volodymyr. "Vziav uchast u konferentsii v Haazi, meta yakoi – poriatunok mizhnarodnoho prava..." ["Participated in a conference in The Hague, the goal of which was to save international law..."]. July 14, 2022. https://t.me/V_Zelenskiy_official/2534. Accessed: July 25, 2023.

Aleksej Tikhonov is a Postdoctoral Researcher in Slavic Linguistics. He earned his doctorate on the 18th century' Czech manuscripts of Protestant refugees in exile in Berlin. He is currently working on his habilitation on the identity-forming function of Slavic languages in German rap. Dr. Tikhonov's research interests include the application of digital humanities methods in philology, corpus linguistics, language contact, and the use of Slavic languages in the online world.

Achim Rabus is a Full Professor of Slavic Philology (Linguistics) and the Managing Director of the Slavic Seminar at the University of Freiburg, Germany. He earned his doctorate with a dissertation on the language of East Slavic spiritual songs in a cultural context and completed his habilitation focusing on the role of language contact in the development of Slavic standard languages. Professor Rabus's research interests encompass Paleoslavistics, the computer-assisted transcription and analysis of Slavic languages using AI methods, and corpus linguistics.