Baltic J. Modern Computing, Vol. 12 (2024), No. 4, 628-635
https://doi.org/10.22364/bjmc.2024.12.4.22

From Manuscripts to Machine-Readable Texts: Developing AI Model for Latvian Autobiographic Heritage

Laura ŠVĪTIŅA

Jāzeps Vītols Latvian Academy of Music

laura.svitina@jvlma.lv

ORCID 0009-0002-0871-4932

Abstract. This article documents a successful collaboration in digitizing historical sources and developing an AI text recognition model to generate automatic transcriptions. Kārlis Paucītis (1891–1967), a Latvian musician, opera clarinetist, bandmaster, and music journalist, chronicled Latvia's 20th-century musical life in his diaries, which have remained largely unresearched despite their historical value. To address this, "Excerpts from Diaries on Musical Life" have been digitized and published online through collaboration with the Archives of Latvian Folklore and the National Library of Latvia. To further advance the accessibility of these historical sources, a custom Handwritten Text Recognition model was developed using the Transkribus platform, achieving 95% accuracy after ongoing refinement and expansion of the training set. This article advocates for the development of a generic Latvian HTR model to broaden accessibility for researchers working with Latvian-language documents. It also emphasizes the critical role of interdisciplinarity, demonstrating how the integration of historical expertise and digital tools can lead to precise and impactful research outcomes.

Keywords: accessibility, digitalization, handwritten text recognition, Transkribus, autobiographic heritage, musician's diaries, Kārlis Paucītis, Latvian music history.

1. Introduction

Recent advancements in handwritten text recognition (HTR) are further advancing the accessibility of historical documents by enabling accurate transcriptions, making texts machine-readable and searchable, and therefore marking an important turning point in the ongoing mass digitization of historical sources. However, limited financial and human resources often lead to the prioritization of specific collections for digitization (Terras, 2022). In Latvia, the efforts to digitize autobiographic heritage are further complicated by varied organizational principles governing life-writing archives (Reinsone et al., 2025). Furthermore, while platforms like Transkribus offer publicly available HTR models for various languages and time periods, no such model is

currently accessible for the Latvian language. This article addresses these challenges by detailing a collaboration to prioritize and digitize Latvian autobiographic heritage, focusing on the musician's Kārlis Paucītis collection. Faced with the limitations of crowdsourcing and the labor-intensive nature of manual transcription, the aim shifts towards developing a custom HTR model using Kārlis Paucītis's diary excerpts as a case study.

2. Historical Sources: Collection of Musician Kārlis Paucītis

In researching music journals and musician organizations during the first period of the Republic of Latvia, Kārlis Paucītis (1891–1967) emerged as a significant figure (Švītiņa, 2021). He devoted his life to music, working as a clarinetist in the Latvian National Opera, a bandmaster, and a music journalist and editor. Despite his contributions, he remains largely forgotten and his legacy still awaits a comprehensive evaluation. The National Library of Latvia (NLL) holds his collection (RXA117). Organized into 132 folders, this collection includes personal documents, professional records, extensive diaries and excerpts from diaries, correspondence, photographs, and various music-related materials, offering detailed insights into Paucītis's work and the broader landscape of 20th-century Latvian musical life, as well as cultural and musical exchanges between Latvia and other countries, particularly in music journalism.

Throughout his life, Paucītis kept a habit of writing down notes and observations. His collection in NLL includes approximately 200 notebooks labeled "diaries" (RXA117,31–53), and at least 17 notebooks called "Excerpts from diaries" (RXA117,23–30;55–63), although determining the total volume is challenging due to the lack of precise metadata (sometimes only the number of notebooks or volumes is listed, while in other cases, only the page count is provided). These excerpts were compiled by Paucītis himself, possibly in the 1960s while living and working in Cēsis under Soviet occupation. They offer insightful commentary and details on the newly formed opera orchestra, vivid descriptions of concerts and composers, and persistent efforts to establish and sustain music journals in Latvia, revealing his experience and collaborations with foreign music journal editors.

As the excerpts were created by the author himself, intimate and personal details present in the diaries were intentionally excluded. Although the excerpts are over 100 years old with all individuals mentioned deceased, this source, while not containing offensive language, should be interpreted within its historical context.

Despite their significance as a primary source, both diaries and excerpts, have remained unresearched. Possibly because of a lack of awareness of the collection's existence in the NLL Department of Rare Books and Manuscripts, the absence of digitally accessible manuscript copies, or Paucītis's intricate handwriting which makes the reading more complicated.

3. Personal Initiative and Institutional Collaboration for Digital Accessibility

In the process of reviewing the excerpts from Kārlis Paucītis's diaries, I quickly recognized their potential value for research on Latvian music history. Around the same

time, I became aware of an ongoing project by the Archives of Latvian Folklore (ALF), part of the Institute of Literature, Folklore, and Art at the University of Latvia, to create an Autobiography Collection. According to their website, ALF invites individuals and institutions to collaborate by submitting materials which are then published online in the Digital Archives of Latvian Folklore website garamantas.lv (Autobiogrāfiju krājums, no date). This platform also supports a crowdsourcing campaign for manuscript transcription and has built an active community of manuscript transcribers. Until the year of 2022 providing 180,000 manually transcribed pages of different collections in machine-readable text (Reinsone and Laime, 2022, p. 60).

Recognizing the opportunity to make Paucītis's work more accessible, I contacted the ALF to explore the possibility of digitizing and publishing these manuscripts online. Since the collection is owned by the NLL, institutional collaboration and coordination among various departments within NLL was necessary. Although there were potential logistical challenges, within a month, ALF formally requested the NLL to prioritize the digitization of selected Paucītis's manuscripts. The request was approved, and the digitization process began shortly after.

Nine notebooks from Paucītis's collection (RXA117,55–RXA117,63), totaling 409 pages from the years 1911 to 1926, were digitized and published online under the title "Excerpts from Diaries on Musical Life".

Despite the active user engagement and a community of transcribers, only one person besides myself attempted to transcribe Paucītis's manuscripts. Upon closer examination, it appears that this person only transcribed pages containing easily readable text (digitized pages with metadata of the physical notebooks). No one else tried to transcribe excerpts from diaries, likely due to the complex handwriting, dense text, and contextual knowledge required. The slow and laborious nature of manual transcription led to the search for a more efficient solution. The Transkribus platform emerged as a widely used software for handwritten text recognition, promising to "unlock the past with AI" (Transkribus, 2024).

4. Using AI and HTR to Automatically Transcribe Documents within The Transkribus Platform

Handwritten text recognition (HTR) relies on artificial intelligence (AI), specifically supervised machine learning. HTR uses artificial neural networks and deep learning algorithms to predict character sequences in handwritten text (Hodel, 2022). The Transkribus platform primarily uses the PyLaia HTR engine, which combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) architectures to convert document images into machine-readable text (Leifert et al., 2024). Transkribus now also offers Transformer-based supermodels for text recognition, which perform better on mixed materials and multiple languages simultaneously, however, a specialized PyLaia model is still recommended for the best results if the material in homogeneous (Transkribus, 2023).

Accessible without requiring expertise in AI or programming, the Transkribus platform provides advanced text editor, AI text recognition, custom AI model training, field and table recognition, publishing and smart search capabilities. Developed through EU-funded research projects, Transkribus has been maintained and expanded by the cooperative READ-COOP, with ongoing contributions from the user community. There

are several subscription plans, including individual use that give 100 free monthly credits. Moreover, extensive resources, user guides and a help center are freely accessible (Transkribus, 2024).

There are over 200 publicly available models for automatic text recognition, developed either by the Transkribus community or the team. Users can easily test these models on their own sources, and if the accuracy is satisfactory, apply them for faster transcription or use them as a basis for creating a custom model. However, this is not the case for the Latvian language, as there is no public model available. Due to this limitation, building a custom AI text model was the only viable option for testing Transkribus. Additionally, it's important to note that each AI text model is trained on documents with specific scripts, languages, time, and place. To achieve the best possible accuracy, particularly for large, but relatively homogeneous collections like Paucītis's diaries, training a custom model would still be the best solution.

Current research on the HTR recognition focuses on more advanced applications, such as testing different HTR engines, developing AI models trained on uncategorized, multilingual, multi-author collections of data, using text recognition for handwritten tables, charts, mathematical expressions (Capurro et al., 2023; Agrawal, Jagtap and Kantipudi, 2024; Leifert et al., 2024). However, there has been little research on applying automatic HTR to the Latvian language. One study has evaluated the performance of custom HTR models trained on Transkribus, using Collection of Cesvaine Secondary School as training data. While using a smaller than suggested training dataset with two different handwritings, a usable and highly accurate text model was created (Slavika, 2024).

5. Methodology: Building the Custom AI Text Recognition Model

To build a custom AI text model, accurate transcriptions of images, referred to as Ground Truth, were necessary. It is recommended to prepare a training set size of at least 10,000 words for each unique handwriting style (Transkribus, 2023). Each image was first processed using Layout Recognition to automatically detect text regions and lines with the publicly available Universal Lines model. However, some lines in Paucītis's diary excerpts were incorrectly segmented due to larger word gaps, requiring manual adjustments to merge the lines correctly before transcription. To address this time-consuming issue, a custom Layout Recognition model was trained using the manually corrected pages. Although this step costs an additional 0.25 credits per page, it proved beneficial, reducing the need for further manual corrections.

With the long-term objective of publishing a critical edition of these diary excerpts, I focused on achieving the highest possible accuracy and consistency throughout the process of creating manual transcriptions and reviewing automatic transcriptions. When deciphering names of international composers, performers, conductors, as well as ensuring the correct spelling of musical works across various languages, doubtful or illegible parts were cross-checked with digitized press materials (www.periodika.lv), mostly searching for published concert programs and reviews. Knowledge of historical context was critical in resolving complex references.

Two types of textual tagging were used based on the degree of readability: “unclear” for partially legible text, adding possible text versions, and “gap” for text that was completely unreadable. Entire lines containing these tags were omitted in the model training process, to avoid learning of mistakes.

Finally, Ground Truth manual transcriptions were prepared for 32 pages, totaling 15,559 words.

After careful editing on the Transkribus platform, the transcriptions were uploaded to Paucītis’s collection on the Digital Archives of Latvian Folklore website www.garamantas.lv, to provide the accessibility during the work process. Each transcription was reviewed again, following the guidelines that differ from the ones that Transkribus provides (LFK, no date). Illegible text needed to be marked with yellow and text lines needed to be formatted so that each new line begins with a diary entry date.

It is important to note that only the initial custom AI text model relied entirely on manual transcription, while subsequent models utilized a combination of automatic transcription followed by manual editing as described previously.

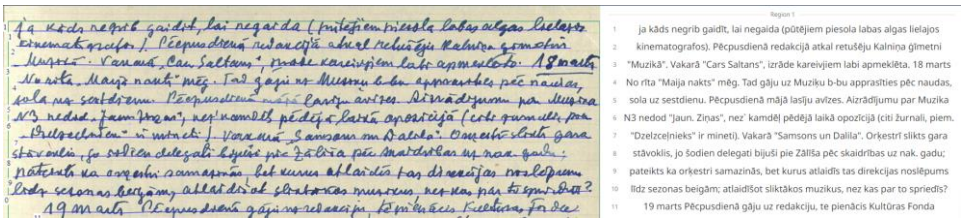

Figure 1. Fragment from Paucītis’s diary excerpt on musical life (NLL RXA117,63; garamantas.lv Ak219-Karlis-Paucitis-03-0035.jpg) alongside Transkribus text editor view

6. Results: Evaluation of the custom AI Text Recognition Models

Typically, the performance of AI text models is evaluated using Character Error Rate (CER) metric, which represents the percentage of incorrectly recognized characters against the total number of characters. A lower CER indicates higher accuracy. According to resources given by Transkribus, models for handwriting are considered highly effective when they achieve a CER between 2% and 8% (Transkribus, 2023).

The initial custom AI model, *Paucitis_1*, was trained on 32 pages, totaling 15,559 words, and achieved a CER of 19.30%. While it provided decent automatic transcriptions and enabled rapid expansion of the training set, it was insufficient for the final version. Paucitis_1 was employed to automatically recognize an additional ten pages of Paucītis’s diary excerpts.

A second model, *Paucitis_2*, was trained on 42 pages, totaling 19,703 words, and showed significant improvement, reducing the CER to 10,10%. The performance increase between the first and second models was primarily due to the larger training set, but growing familiarity with Paucītis’s handwriting and his thought process also allowed me to correct earlier mistakes and clarify words that had previously been tagged as

unclear. Continued refinement and expansion of the training set further improved the accuracy as shown in Table 1.

Table 1. Performance of custom HTR models, showing improvements in CER with increasing training set size

Custom HTR model	*Training Set Size (words)*	*CER (%)*
Paucitis_1	15 559	19,30
Paucitis_2	19 703	10,10
Paucitis_3	29 087	7,60
Paucitis_4	32 968	6,30
Paucitis_5	40 480	5,70
Paucitis_6	48 700	5,30
Paucitis_7	48 700	5,50
Paucitis_8	57 598	4,56

Both models *Paucitis_6* and *Paucitis_7* were trained using the same training set. But a comparison was undertaken with changing advanced settings – training a model with an existing base model and with existing line polygons. Although the line polygons, an area that encases all the handwritten text, were sometimes incorrect, it was unclear whether editing them was necessary if the baselines were accurate. Transkribus manuals caution that models trained using existing line polygons will perform best when applied to documents with similar line polygon structures. Model *Paucitis_6* was trained without using an existing base model but keeping the existing line polygons, while *Paucitis_7* was trained with the preexisting *Paucitis_5* model as a base but without using existing line polygons. The differences in CER between these two models were minimal. From this comparison, it was concluded that the final model will be trained without the existing base model or line polygons to avoid learning from previous errors and to increase the usability of the model.

The latest model, *Paucitis_8*, was trained on 123 pages, totaling 57,598 words, and achieved the lowest CER of 4,56%, demonstrating a high level of accuracy for the Paucītis handwriting.

To further validate the effectiveness of the custom HTR model, a comparison was made between the automatic transcription generated by the model *Paucitis_6* and the manually corrected version of the same page. The model's performance was assessed using the compare tool integrated into Transkribus eXpert desktop software. Evaluation metrics included both CER and Word Error Rate (WER). For one test page, the model achieved a CER of 3,81% and a WER of 14,78%. While the CER on a specific test page was lower than the model's overall CER, the higher WER indicates that character errors were spread across multiple words.

7. Discussion

The results suggest that an iterative training process, combined with an expanding training set, a custom Layout Recognition model, and constant review and cross-checking of historical sources, proved to be an effective strategy for developing a custom

HTR model. While manual editing remains necessary to achieve the highest possible accuracy, the model's performance is sufficient to make Paucītis's manuscripts machine-readable and searchable, therefore advancing accessibility.

For large collection like Paucītis's, a custom model is beneficial, giving improvements in both efficiency and accuracy. Moreover, the potential applications of this custom model extend beyond transcription for the whole collection or the creation of a digital edition. The transcribed manuscript pages can now be explored through various digital research methods. A complete transcription of his diaries, compared to the excerpts, could reveal insights into what Paucītis chose to omit or deemed unimportant, offering a deeper understanding of the relationship between his daily life and work, his attitudes and more personal perspectives.

Although Paucītis's writings remain under copyright protection until 2037, my use of his work exclusively for research purposes does not constitute copyright infringement. However, publishing or sharing these Ground Truth materials for further use would require additional permissions. Considering this, these transcriptions could contribute to developing a generic HTR model for the Latvian language. Such a model could be made publicly available through Transkribus, addressing current limitations for users who wish to test the platform or quickly produce Ground Truth data for new custom models in Latvian. Although a generic model might not reach the same accuracy as a custom one, it would be valuable for projects involving a broad range of documents with diverse handwriting styles (Pinche, 2023).

Additionally, the Transkribus platform has recently introduced new features. These include the integration of Transkribus into existing workflows and an automated text-image alignment feature. Given the extensive amount of images and separate text transcriptions on the Digital Archives of Latvian Folklore, these features could greatly enhance the development of a powerful generic HTR model for the Latvian language.

8. Conclusion

This article highlights the power of coordinated efforts in supporting researchers' initiatives and making valuable historical sources publicly available. The custom HTR model developed for Paucītis's diary excerpts using the Transkribus platform has further advanced the accessibility, demonstrating its usefulness for Latvian language documents.

While custom HTR models deliver the best results for homogeneous materials and are especially suited for large collections like Paucītis's, the development of a generic Latvian HTR model in Transkribus could benefit a broader range of researchers, lowering the barriers to using this powerful digital tool. The process of creating accurate transcriptions underscores the importance of interdisciplinarity, as the integration of historical expertise and the implementation of digital tools enables the production of precise and meaningful research outcomes.

Acknowledgments

This article was supported by the project "Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia" (No. VPP-IZM-DH-2022/1-0002) of the State Research Program "Digital Humanities", funded by the Latvian Council of Science.

References

Agrawal, V., Jagtap, J., Kantipudi, M.V.V.P. (2024). Exploration of advancements in handwritten document recognition techniques, *Intelligent Systems with Applications* **22**, https://doi.org/10.1016/j.iswa.2024.200358.

Autobiogrāfiju krājums [Collection of Autobiographies] (no date). Available at: https://garamantas.lv/lv/repository/1115628/Autobiografiskas-kolekcijas

Capurro, C., Provatorova, V., Kanoulas, E. (2023). Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection, *Heritage* ***6***(12), pp. 7482–7494. Available at: https://doi.org/10.3390/heritage6120392.

Hodel, T. (2022). Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities, *in* L. Jaillant (ed.) *Archives, Access and Artificial Intelligence*, pp. 157–178. Available at: https://www.jstor.org/stable/jj.11425482.9

LFK (no date). Latviešu folkloras krātuves digitālais arhīvs garamantas.lv *Atšifrēšanas vadlīnijas* [Digital Archives of Latvian Folklore garamantas.lv, Suggestions for transcriptions] Available at: https://media.garamantas.lv/files/atsifresanas-vadlinijas.pdf

Leifert, G., Romein, C.A., Rabus A., Ströbel, P.B., Hodel, T. (2024). *Transkribus and Beyond: Pioneering the Future of Transcription Technology*, Poster presentation. Transkribus User Conference '24, knaw.nl

Pinche, A. (2023). Generic HTR Models for Medieval Manuscripts. The CREMMALab Project, *Journal of Data Mining & Digital Humanities, Historical Documents and automatic text recognition*. Available at: https://doi.org/10.46298/jdmdh.10252.

Reinsone, S. and Laime, S. (2022). Latviešu folkloras krātuves digitālais arhīvs garamantas.lv: priekšvēsture un attīstība, Letonica 47, pp. 52–69.

Reinsone, S., Ļaksa-Timinska, I., and Žvarte, E. (Forthcoming 2025). Beyond the Archive Shelves: Navigating Digital Archiving of Life Writing Legacy through Public Involvement. *a/b Autobiography* [2025]

Slavika, D. L. (2024). *Leveraging Artificial Intelligence Solutions for the Transcription of Handwritten Folklore Manuscripts in Latvian.* Master thesis. RTU.

Švītiņa, L. (2021). Latvijas Mūziķu biedrība (1922–1939): darbības profils un koncertakcija "Mūziķu diena" [Latvian Musicians` Society (1922–1939): Activity Profile and Concert Campaign "Musicians Day"]. *Mūzikas akadēmijas raksti* XVIII, pp. 39–74.

Terras, M. (2022). Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription, *in* L. Jaillant (ed.) *Archives, Access and Artificial Intelligence*, pp. 179–204. Available at: https://www.jstor.org/stable/jj.11425482.10.

Transkribus (2023). *Training Models. Data Preparation.; Text Recognition. Super Models.* Available at: https://help.transkribus.org/data-preparation, https://help.transkribus.org/super-models

Transkribus (2024). *Transkribus*. Available at: https://www.transkribus.org/

Received December 1, 2024, accepted December 10, 2024