Automated Character Recognition in Early Tibetan Canonical Manuscripts

Eric Werner & Markus Viehbeck
(TMPV, University of Vienna)

The use of computational methods is advancing rapidly across all domains of natural sciences, social sciences, and humanities. Tibetan studies, despite its seemingly niche focus, has been actively embracing this trend. Given the significance of textual studies to the field, advancements in this area are particularly prominent and diverse.[1] These include work on automated character recognition of the contents of Tibetan written artifacts, which is an essential step in producing larger corpora of searchable Tibetan e-texts for any kind of subsequent computational analysis. In the past, a more widely known implementation of Tibetan optical character recognition (OCR) was developed under the name Namsel OCR,[2] and Google's built-in OCR feature in Google Drive/Google Docs has also been (and continues to be) widely used. While these earlier tools were primarily developed to recognize Tibetan *dbu can* script in printed materials, more recent work has included manuscripts as well. Of particular note are the Handwritten Text Recognition (HTR) models developed within the TibSchol project at the Austrian Academy of Sciences using the Transkribus platform, which are the first to address Tibetan *dbu med* script.[3] Another Transkribus model, focusing on printed Tibetan newspapers, has been published by the Divergent Discourses project.[4] More recently, collaborative, large-scale efforts to develop Tibetan OCR/HTR for various scripts and media have been supported by the Buddhist Digital Resource Center and included several other related initiatives.[5]

---

1 The edited volume by Meelen, Hill, and Faggionato 2024 showcases several important avenues.

2 See Rowinski and Keutzer 2016. This also summarizes earlier attempts in Tibetan OCR.

3 See Griffiths 2024.

4 Developed by Franz Xaver Erhard as Tibetan Modern U-chen Print, accessible on Transkribus: https://app.transkribus.org/sites/uchan/about; accessed Nov. 14. 2024.

5 As a result of these collaborative efforts, BDRC has recently released an OCR app with a wide range of applications: https://www.bdrc.io/blog/2025/03/14/bdrc-announces-the-release-of-ocr-app-for-tibetan/; accessed Jun. 30. 2025.

Eric Werner & Markus Viehbeck, "Automated Character Recognition in Early Tibetan Canonical Manuscripts", *Revue d'Etudes Tibétaines*, no. 79, Octobre 2025, pp. 55-65.

New Findings of Early Tibetan Canonical Manuscripts: The Namgyal Sūtra Collection

As a long-term research initiative, the Tibetan Manuscript Project Vienna (TMPV) specializes in the documentation and research of Tibetan canonical literature and, with its Resources for Kanjur and Tanjur Studies (rKTs) archive, hosts the largest online database dedicated to Tibetan canonical sources. Through the activities of the TMPV, several early canonical manuscript collections have been documented. Of particular importance are the so-called "Sūtra collections" (*mdo sde*), an early Tibetan corpus of approximately 435 canonical texts arranged in a specific order in thirty volumes. Several such Sūtra collections have recently been documented in Dolpo and Mustang in the Nepal-Tibet border region.[6] These include a manuscript collection from Namgyal (*rnam rgyal*) Monastery in Mustang, documented in collaboration with the SOAS-based project Tibetan Buddhist Monastery Collections Today (directed by Christian Luczanits). This fragmented Sūtra collection was not only the first to be comprehensively documented, but its manuscripts are also of outstanding artistic quality in terms of material support, scribal work, and the illuminations that adorn the beginning and end of the volumes. Their codicological features, such as layout and ornamentation as well as orthography and paleography, indicate a rather early date.[7] While the early fourteenth century has been proposed as an initial tentative estimate for its production,[8] more recent comparative research with data from related collections in Dolpo, including C14 analysis of paper samples, suggests an earlier date for the Namgyal Sūtra collection as well, potentially as early as the twelfth century.[9] In addition to its early age for such a large collection of what we can now call canonical texts, the structural features and textual contents of this material make it of considerable interest for further philological research, but work on these manuscripts also poses certain challenges.

---

6 The manuscript features of these local collections are discussed in detail in Viehbeck, forthcoming. The documentation activities are also reported on the TMPV website, under Documentation: https://tmpv.univie.ac.at/; accessed Nov. 14. 2024.

7 The textual contents of this collection, along with the images, were published at rKTs: http://www.rkts.org/collections.php?id=1Ng; accessed Nov. 15. 2024. Additionally, these were shared with BDRC: https://library.bdrc.io/show/bdr:MW2KG229028; accessed Nov. 15. 2024.

8 See Luczanits and Viehbeck 2021, 42–43, 142–143.

9 These issues of comparative dating are discussed in detail in Viehbeck, forthcoming.

Challenges and Research Objectives

As outlined in earlier research,[10] the Sūtra collection represents a canonical model that evolved prior to and alongside the model of structured Bka' 'gyurs, which became dominant with its emergence in the early fourteenth century. The differences between these two models relate not only to major structural variations in the arrangement of canonical texts that are already known, but there are also instances in which a Sūtra collection encompasses several texts that are entirely absent from subsequent structured Bka' 'gyurs. On a finer scale, it is apparent that even in cases of commonly identified texts, the version found in the Sūtra collection diverges from versions found in other sources. However, the extent of these differences varies considerably. As a working hypothesis, it seems plausible to assume that some can be attributed to the specific process of production and transmission, i.e., repeated manual copying, while others point to divergent sources and translations, i.e., the use of an entirely different source text for the translation, or a modified translation of the same source.

In order to address this broader set of questions regarding the production and transmission of alternative canonical sources by using computational methods and on a larger scale, the availability of an e-text of this collection is an obvious *desideratum*. The Namgyal OCR model was therefore developed primarily as a first step in addressing these research questions. Furthermore, the model should prove beneficial to others working with similar sources.

While models for Tibetan printed *dbu can* script and several types of handwritten *dbu med* script are currently available, a dedicated model for handwritten *dbu can* script was still missing. Moreover, as these manuscripts date to such an early period, they exhibit some distinctive features. For example, in terms of layout and line recognition, the presence of decorative string hole markings in the center of the folios has presented a major challenge. This is equally true for typical features of archaic orthography (such as *ma ya btags*, *da drag*, *gi gu log*, *'a* suffix, *anusvāra*) and paleographic features such as the use of horizontal ligatures as well as the specific shapes of individual characters. The Namgyal model should therefore be of direct utility to anyone working with similar early Tibetan *dbu can* handwritten material. Furthermore, its development has contributed to the advancement of wide-scope OCR models for Tibetan, as described above. The following workflow was used to train this model.

[10] See Viehbeck 2020, Luczanits and Viehbeck 2021, 341–361.

The OCR Pipeline

To digitize the collection, we opted for two-stage approach comprising line, viz. layout detection, via image segmentation inspired by Grüning *et al*. 2019. Due to language-specific features, such as long descending letters and complex letter stacks that occupy significant vertical space, Tibetan manuscripts (and block prints) typically exhibit minimal interlinear spacing, if any, and frequently display instances of characters touching adjacent lines (see examples below).

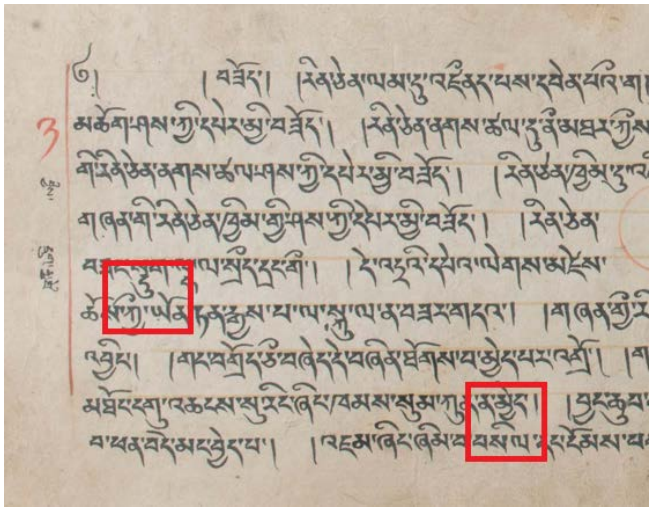


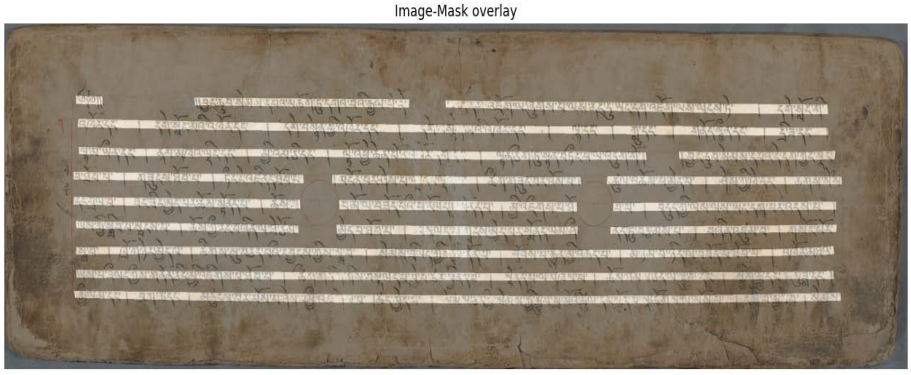
Fig. 1: Examples of characters touching adjacent lines

Therefore, we decided to train a model for line detection, as opposed to determining the interlinear space, and to employ classical post-processing methods using OpenCV for line extraction (including rotation correction, dilation, masking etc.) to extract the final lines. After several experiments with different backbone architectures commonly used in image segmentation tasks (such as U-NET), DeepLabv3+[11] in combination with Dice and Focal-Loss proved to be the most promising approach to handle background exclusion most effectively for the given use case.

The prepared dataset comprises 100 annotated images with annotations for images, lines, and marginalia (see examples below). Since Tibetan manuscripts can have significantly discontinuous text segments per reading line (i.e., with substantial gaps between the text segments), we experimented with two distinct versions of the dataset

[11] See Chen *et al*. 2018.

to determine whether these gaps necessitate consideration under all circumstances.



Figs. 2 & 3: The same folio being annotated with and without considering gaps in the line

Since the entire collection maintains a relatively consistent page layout and resolution, it proved feasible to train a model based on a down-sampled version of the entire image, rather than employing the more common approach of image tiling. Accordingly, we down-sampled the entire image to a resolution of 1024 x 320 pixels and trained two image segmentation models to identify different layout elements—primarily lines, but also images, image captions, and marginalia (see figures below). We used Transkribus as an annotation tool and employed a custom export pipeline to prepare the training data. The models were then trained using a custom training pipeline built on PyTorch and PyTorch SegmentationModels.[12] The datasets and annotations are available on Zenodo and HuggingFace, including the

[12] https://github.com/qubvel-org/segmentation_models.pytorch.

original annotations in PageXML format as well as the exported masks for layout elements and lines-only.[13]

Model	Train/Val/Test Samples	Test Dice-Score	Test Jaccard-Index
Layout	80 / 10 / 10	0.95	0.64
Lines	80 / 10 / 10	0.88	0.79

Table 1: Summary of trained segmentation models

Fig. 4: Example highlighting the detected elements: images, image captions, margins, lines (each line numbered), and the determined line breaks (blue circles)

Fig. 5: Example of a multi-class mask used for image segmentation

[13] All data generated for the project is available on Zenodo: https://zenodo.org/records/1424773; further also via HuggingFace: https://huggingface.co/datasets/Eric-23xd/GlomanThang-ImageSegmentation-sensitive, https://huggingface.co/datasets/Eric-23xd/GlomanThang-ImageSegmentation.

Fig. 6: Overview of the images used in the dataset

Training domain-specific OCR/HTR models remains challenging and working with datasets in under-resourced languages often requires dedicated research to enhance performance in these areas. The OCR dataset for this project comprises 508 pages drawn from different volumes, with an emphasis on covering the variety of slight script variations across the collection as well as the full orthographic range, particularly the more complex and uncommon Tibetan syllables often found in dhāraṇīs or tantric texts. The dataset was prepared using the (now deprecated) Transkribus offline client, and a custom export pipeline was written in Python to generate a line-based dataset for training custom OCR models using the Tensorflow and, subsequently, the PyTorch library. For this project, we adopted Easter2 ("Easy and Scalable Text Recognizer") by Chaudhary *et al*. 2022, which leverages 1D convolutions in combination with residual blocks and CTC loss to facilitate OCR training. Due to its competitive performance on small datasets and lightweight architecture, it proved to be the architecture of choice for the given scenario and resource constraints. Following initial tests based on the original implementation provided in Keras and Tensorflow 2,[14] we ported the architecture to PyTorch and added custom image augmentations via albumentations (shearing, embossing, pixel dropout, Gaussian noise, etc.).

[14] https://github.com/kartikgill/Easter2.

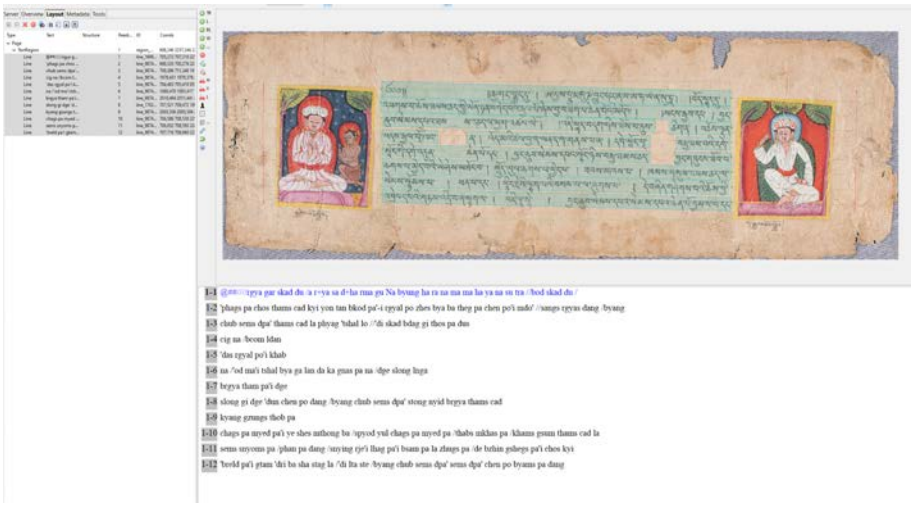
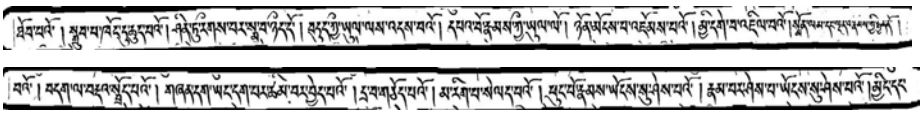


Fig. 7: Dataset preparation in Transkribus

In the course of an annotation process commenced in summer 2022, the following dataset—hereafter referred to as Namgyal dataset—was generated:

Training Samples: 4107
Validation Samples: 411
Test Samples: 644

Each line was resized and padded to 3200x100 pixels. This increased image height (64 pixels being a more common choice) was chosen to mitigate potential information loss from previous down-sampling, given that Tibetan letter stacks are much more complex and detailed than those in Roman scripts.



Figs. 8 & 9: Line image samples extracted from Transkribus

Trained OCR Models

The following overview summarizes the models trained during the course of the project. We generally refer to the models trained on the Namgyal dataset as "Early Tibetan Manuscript Uchan." For

comparison, a Transkribus model was trained using the PyLaia backend. The other models are based on our PyTorch implementation of the Easter2 architecture using Wylie-encoded labels. We trained models solely on the Namgyal dataset and also fine-tuned BDRC's latest BigUchan model using the same dataset. The trained models can be used via the inference pipeline provided on Github[15] or as part of BDRC's desktop application, forthcoming in 2025. The Transkribus model is accessible through the Transkribus ecosystem under the name Early Tibetan Manuscript Uchan. An evaluation of the test set using the BigUchan model without fine-tuning is also provided.[16]

Model	Architecture	Encoding	CER / Accuracy on Validation Set
Early Tibetan Manuscript Uchan	PyLaia HTR (Transkribus)	Wylie	1.9% (=0.019?)[17]
Early Tibetan Manuscript Uchan	Easter2	Wylie	CER: 0.041
Early Tibetan Manuscript Uchan	CRNN	Wylie	CER: 0.04
BDRC's BigUchan 1	Easter2	Wylie	CER: 0.069
Fine-tuned Namgyal model on BigUchan 1	Easter2	Wylie	CER: 0.05

Table 2: Overview of trained models and the mean CER on the test set

[15] The code is accessible on https://github.com/eric86y/Namgyal-OCR. The repository additionally includes a Jupyter notebook that demonstrates a sample workflow. However, some technical setup, including a Python environment and requisite packages, is necessary.

[16] The trained models are available on Huggingface. For the Easter2 version, see https://huggingface.co/Eric-23xd/EarlyTibetan-Manuscript-Uchan. For the CRNN version, see: https://huggingface.co/Eric-23xd/EarlyTibetan-Manuscript-Uchan-CRNN. For the version fine-tuned on BigUchan 1, see https://huggingface.co/Eric-23xd/EarlyTibetan-Manuscript-Uchan-BigUchan. All models are also available on the project's Zenodo page: https://zenodo.org/records/14247731. The official Github repository of the project can be accessed here: https://github.com/eric86y/Namgyal-OCR.

[17] Transkribus provides CER scores for both training and validation sets. However, it remains unclear whether the validation set is actually used as a proper test set. To facilitate a meaningful performance comparison, we used the Transkribus's "validation data" as hold-out test set.

Acknowledgements

The research presented in this article was conducted as part of the "Himalayan Sūtra Collections" project (P35697), funded by the Austrian Science Fund (FWF). The article was proofread by Filippo Brambilla.

References

Chaudhary, Kartik, and Raghav Bali. 2022. "Easter2.0: Improving Convolutional Models for Handwritten Text Recognition." https://arxiv.org/abs/2205.14879

Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." https://arxiv.org/abs/1802.02611.

Griffiths, Rachael. 2024. "Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script." *Revue d'Etudes Tibétaines* 72: 43–51. (doi:10.1553/TibSchol_ERC_HTR)

Grüning, Tobias, Gundram Leifert, Tobias Strauß, Johannes Michael, and Roger Labahn. 2019. "A Two-stage Method for Text Line Detection in Historical Documents." *International Journal on Document Analysis and Recognition* 22, 285–302. (https://doi-org.uaccess.univie.ac.at/10.1007/s10032-019-00332-1)

Luczanits, Christian, and Markus Viehbeck. 2021. *Two Illuminated Text Collections of Namgyal Monastery: A Study of Early Buddhist Art and Literature in Mustang*. Vajra Academic Vol. I. Kathmandu: Vajra Books.

Meelen, Marieke, Nathan Hill, and Christian Faggionato (eds.). 2024. *Proceedings of the IATS 2022 Panel on Tibetan Digital Humanities and Natural Language Processing*. *Revue d'Etudes Tibétaines* 72.

Rowinski, Zach, and Kurt Keutzer. 2016. "Namsel: An Optical Character Recognition System for Tibetan Text." *Himalayan Linguistics* 15(1): 12–30. (http://dx.doi.org/10.5070/H915129937)

Viehbeck, Markus. forthcoming. "Sūtra Collections in the Himalayan Borderlands: Local Manuscripts of an Early Tibetan Canonical Model." In *Canons, Kangyurs, and Collections—Multidisciplinary Approaches in the Study of Tibetan Canonical Literature*, edited by Markus

Viehbeck, Filippo Brambilla, and Kurt Tropper. Vienna: Austrian Academy of Sciences Press.

Viehbeck, Markus. 2020. "From Sūtra Collections to Kanjurs: Tracing a Network of Buddhist Canonical Literature across the Western and Central Himalayas." *Revue d'Etudes Tibétaines* 54: 241–260.

❖