Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script[*]

Rachael M. Griffiths
(Austrian Academy of Sciences)

The use of advanced computational methods for the analysis of digitised texts is becoming increasingly popular in humanities and social science research. One such technology is Handwritten Text Recognition (HTR), which generates transcripts from digitised texts with machine learning approaches, to enable full-text search and analysis. Up to now, HTR models for Tibetan manuscripts in cursive script have not been available. This paper introduces work carried out as part of the *The Dawn of Tibetan Buddhist Scholasticism (11th-13th)* (TibSchol) project at the Austrian Academy of Sciences, which is utilising the Transkribus platform to explore possible solutions to automate the transcription of Tibetan cursive scripts. It presents our methodology and preliminary results along with a discussion of the limitations and potential of our current models.

1. *Introduction*

Handwritten Text Recognition (HTR) is an active research field that has developed significantly over the last decade, making great strides in its ability to automatically transcribe texts, especially those in Roman script (Nockels 2022). Focus now is being applied to extending this to other scripts–including Devanagari (Merkel-Hilf 2022), Hebrew (Digitizing Jewish Studies (DiJeSt) 2020), and Pracalit script (O'Neill & Hill 2022)–and offers great potential to those studying texts in a wider range of languages. In the context of Tibetan, several initiatives have been undertaken to develop Optical Character Recognition (OCR) systems. Notable OCR implementations for

[*] Rachael Griffiths, "Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script", *Revue d'Etudes Tibétaines*, no. 72, Juillet 2024, pp. 43-51.

The writing of this paper was facilitated through the project *The Dawn of Tibetan Buddhist Scholasticism (11th-13th)* (TibSchol). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001002 TibSchol). The results presented are solely within the author's responsibility and do not necessarily reflect the opinion of the European Research Council or the European Commission who must not be held responsible for either contents or their further use.

Tibetan include Namsel OCR[1] and Google Drive/Google Docs. Additionally, projects and organisations such as the Buddhist Digital Research Centre (BDRC, https://www.bdrc.io) and Esukhia (https://github.com/Esukhia/) are actively engaged in research and development efforts in OCR and HTR. However, despite these endeavors, publicly available HTR models for Tibetan are currently unavailable.

Expanding the abilities of HTR models to Tibetan manuscripts is one strand that is being explored as part of the ERC-funded project *The Dawn of Tibetan Buddhist Scholasticism (11th–13th)* (TibSchol) at the Austrian Academy of Sciences. The project is carrying out an extensive study of the formative phase of Tibetan Buddhist scholasticism, utilising a large number of recently surfaced works. Notably it explores texts that were published as manuscript facsimile in the *bKa’ gdams gsung ’bum* (*Collected Works of the Kadampas*) (dPal brtsegs bod yig dpe rnying zhib ’jug khang 2006-2015). Additional relevant manuscript sources are accessible online via BDRC. As TibSchol is a text-based project, HTR offers the possibility of facilitating full-text mining and analysis for a broad-scale approach to the corpus by relying on machine-readable transcriptions of these sources.

2. *Method*

For this task, we have been using Transkribus, a popular platform for transcribing, annotating, and searching historical manuscripts, which can be run on a local machine or in an online web interface (Kahle, Colutto, Hackl & Mühlberger 2017). It allows users to train text recognition models based on images of handwritten text that are lined up with corresponding diplomatic transcriptions which are called ‘ground truth‘. It also provides pre-trained HTR models in a range of scripts, although no public model is currently available in any Tibetan script. We tested a private model trained by Esukhia for handwritten *uchen* (Tib. *dbu can*, ‘headed script’), however it was unable to transcribe the *ume* (Tib. *dbu med*, ‘headless script’) works in our corpus. As such, we have had to train a new HTR model from scratch.

As a guideline for creating an HTR model, Transkribus recommends preparing 5,000-15,000 words (25–75 pages) of transcribed material. In general, the more training data used, the higher the accuracy of the HTR model will be. The metric used to assess the accuracy of an HTR model in Transkribus is Character Error Rate (CER), that is the percentage of character-level errors in the recognised text compared to the ground truth text. A CER under 10% is considered efficient for automatic transcription, however, to maximise the

[1] https://github.com/zmr/namsel

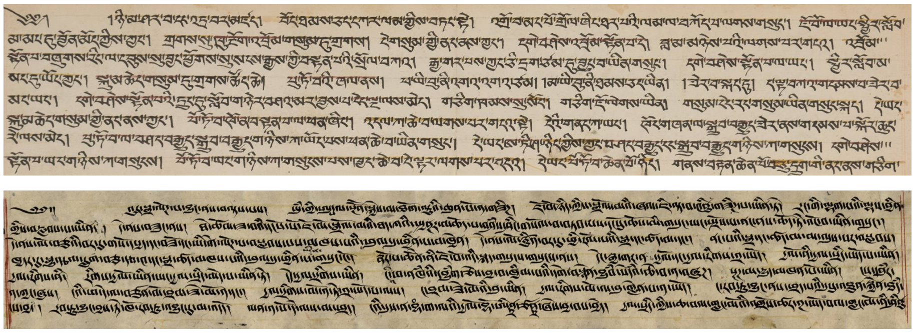


Fig. 1 – Examples of handwritten uchen (top, BDRC W3CN2815) and ume (bottom, BDRC W1NLM1277).

usability of transcribed texts for the text mining and analysis that will form the bedrock of our project, we are aiming for a CER of 5% or lower.

Our workflow began with making fundamental decisions about which scripts and texts the HTR will be required to read and accordingly, which images and transcripts will form a suitable ground truth. The manuscripts in our corpus are written in a variety of scripts (Fig. 2), which makes the possibility of training a general model more challenging. A further challenge is the quality of the images, which varies considerably throughout the corpus, and many folios contain interlinear and/or marginal insertions that are difficult to read. As such, we decided to begin with training a script-specific model, choosing *drutsa* (Tib.*’bru tsha*) due to the quality of images available in this script. We selected five manuscripts (totalling approximately 300 folios and 2500 lines) in *drutsa* script for training that were clearly legible and for which we already had a transcription in the so-called ”Wylie” system, which uses Roman characters, without diacritics, to render univocal combinations of letters in Tibetan syllables (Wylie 1959).

The next step was running the Layout Analysis (LA) tool on manuscript images imported into Transkribus. This tool, which is integrated into the platform, automatically analyses the structure and layout of a document to identify its different components such as text regions, images, and other elements. Before it can transcribe, the HTR model must be able to clearly define what it is seeking to transcribe. The default LA tool did not generate accurate results on the Tibetan manuscripts we had selected for training; results often included multiple text regions across one folio, the omission of lines or lines only being partially recognised, and stains or markings on the manuscript

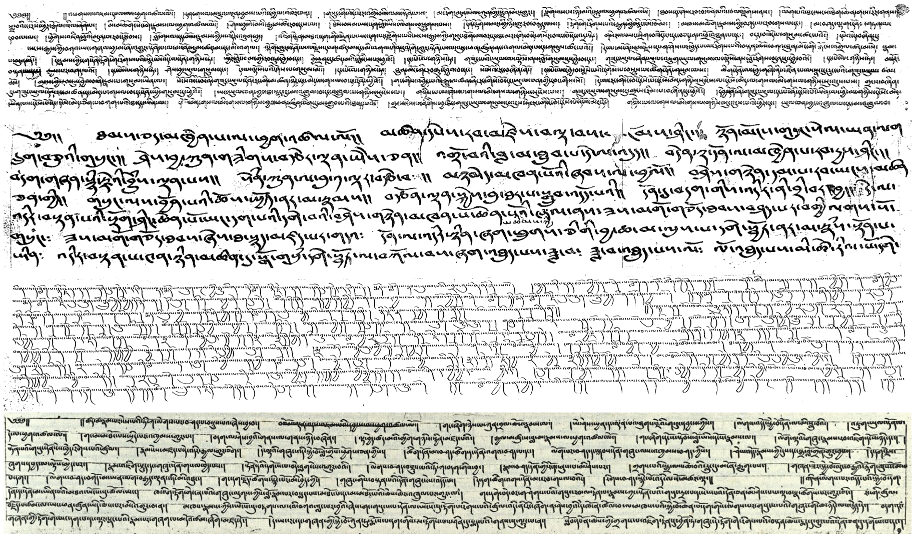


Fig. 2 – Examples of scripts found in the TibSchol corpus (top to bottom, BDRC W26453, BDRC W1CZ1224, BDRC W12170, BDRC W1KG12371).

marked as a text. The number of errors were significant enough that manual correction was not a feasible option (see first image in Fig. 4).

Initially, improvement in the LA was achieved through pre-processing the images before uploading them to Transkribus. We used OpenCV (https://www.opencv.org) to affect image sharpness, resolution, and noise using the filter2D and fastNlMeansDenoising functions (Fig. 3). These image enhancements improved the results of the default LA tool, although it continued to segment stains and markings as baselines, which required manual deletion. We also trialed a Python pre-processing script developed by Esukhia, which automatically generates baselines that stretch across an entire text region (https://github.com/Esukhia/custom-script-for-transkribus). The baselines created are straight lines, which unfortunately are not compatible with our images, some of which are warped and contain marginalia and interlinear additions. In June 2022, Transkribus launched a new feature, where users can train a baseline model specific to their document typology (see Transkribus (2022)). We tested this tool to see if it could train a customised baseline model based on examples from our corpus.

The baseline model was trained on 160 folios that had been manually segmented, keeping aside 10% as a Validation Set (a Validation Set is used to evaluate the model’s performance on unseen data during the training process). A

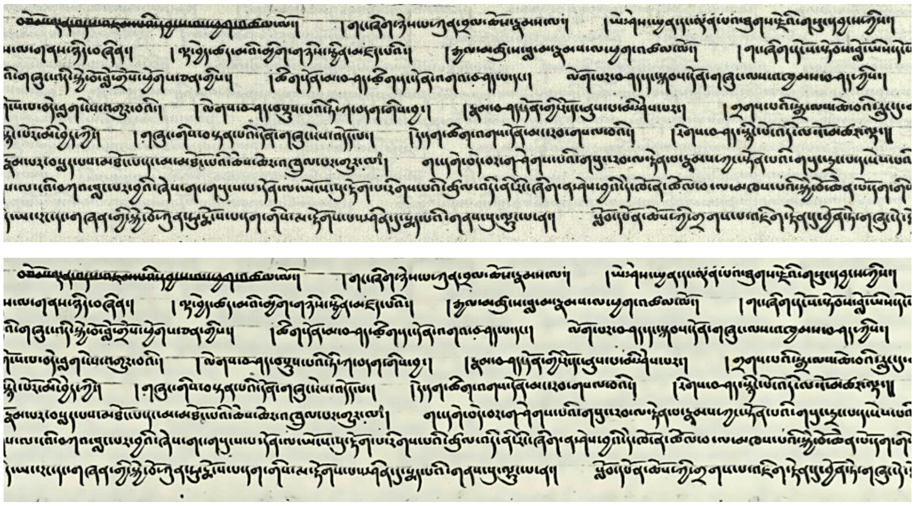


Fig. 3 – Section of folio from Tshad ma rnam par nges pa'i Ti ka legs bshad bsdus pa (BDRC W1KG12371) before (top) and after (below) pre-processing using OpenCV.

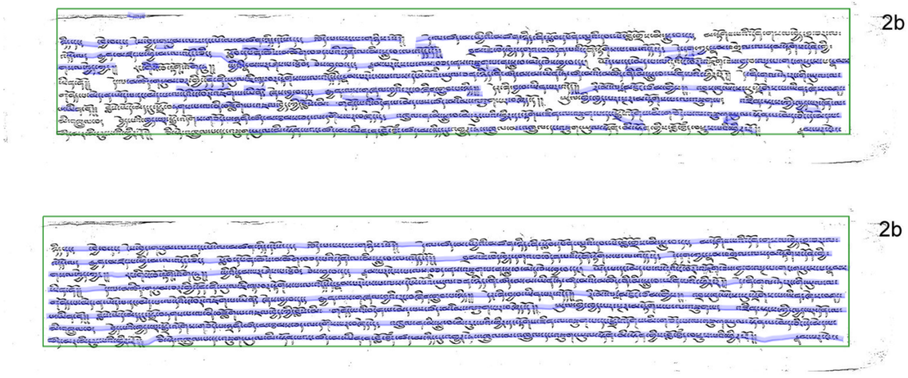
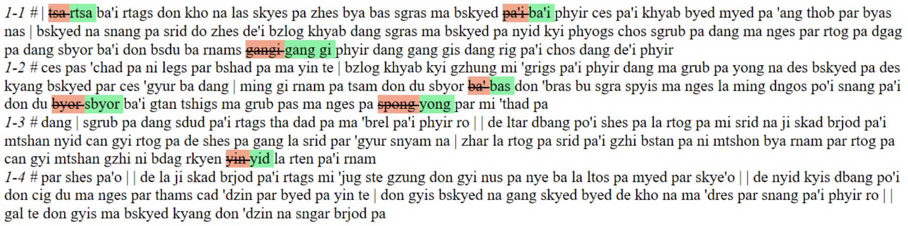


Fig. 4 – Results of the default LA tool (top) and custom baseline model (bottom) on the same folio from Tshad ma'i bstan bcos sde bdun rgyan gyi me tog (BDRC W00KG03838).

Loss on the Training Set (LTS) and Loss on the Validation Set (LVS) below 10% allows for an effective automatic segmentation. The TibSchol baseline model has an LTS of 4.3% and LVS of 3.8%. As can be seen in Fig. 4, the model still requires some manual correction, although this is minimal. As we were satisfied with the results, the baseline model was then run on the remaining folios uploaded to Transkribus.

Once the layout was finalised, the transcriptions were added to the document editor and HTR training could begin. When we first started working with Trankribus, it hosted two HTR engines: HTR+, developed by the CITlab of the University of Rostock, and PyLaia, a PyTorch-based model developed by the Technical University Valencia. During initial tests, PyLaia appeared to struggle with reading ’curved’ lines (see the results of Model 5 in Fig. 6)–some of our images are warped–while the HTR+ engine produced better results. Thus we decided to continue using HTR+ for training.

To more accurately gauge a model’s performance on the Validation Set, it is possible to compare the ground truth of a page with the transcription produced by the model using ’Compare Text Version’ (Fig. 5). This not only flags sticking points for the model but also errors in our ground truth, which were then manually corrected before training a new model with more data.



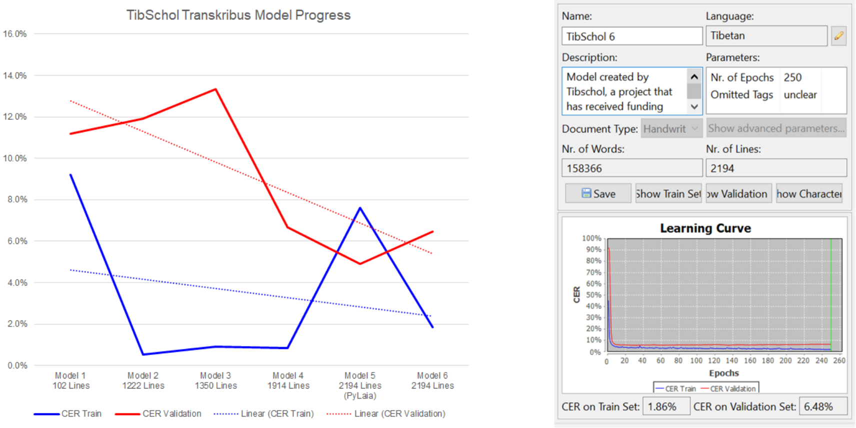
1-1 # | ~~tsa~~ rtsa ba'i rtags don kho na las skyes pa zhes bya bas sgras ma bskyed ~~pa'i~~ ba'i phyir ces pa'i khyab byed myed pa 'ang thob par byas nas | bskyed na snang pa srid do zhes de'i bzlog khyab dang sgras ma bskyed pa nyid kyi phyogs chos sgrub pa dang ma nges par rtog pa dgag pa dang sbyor ba'i don bsdu ba rnams ~~gangi~~ gang gi phyir dang gang gis dang rig pa'i chos dang de'i phyir
1-2 # ces pas 'chad pa ni legs par bshad pa ma yin te | bzlog khyab kyi gzhung mi 'grigs pa'i phyir dang ma grub pa yong na des bskyed pa des kyang bskyed par ces 'gyur ba dang | ming gi rnam pa tsam don du sbyor ~~ba'~~ bas don 'bras bu sgra spyis ma nges la ming dngos po'i snang pa'i don du ~~byor~~ sbyor ba'i gtan tshigs ma grub pas ma nges pa ~~spong~~ yong par mi 'thad pa
1-3 # dang | sgrub pa dang sdud pa'i rtags tha dad pa ma 'brel pa'i phyir ro | | de ltar dbang po'i shes pa la rtog pa mi srid na ji skad brjod pa'i mtshan nyid can gyi rtog pa de shes pa gang la srid par 'gyur snyam na | zhar la rtog pa srid pa'i gzhi bstan pa ni mtshon bya rnam par rtog pa can gyi mtshan gzhi ni bdag rkyen ~~yin~~ yid la rten pa'i rnam
1-4 # par shes pa'o | | de la ji skad brjod pa'i rtags mi 'jug ste gzung don gyi nus pa nye ba la ltos pa myed par skye'o | | de nyid kyis dbang po'i don cig du ma nges par thams cad 'dzin par byed pa yin te | don gyis bskyed na gang skyed byed de kho na ma 'dres par snang pa'i phyir ro | | gal te don gyis ma bskyed kyang don 'dzin na sngar brjod pa

Fig. 5 – Errors in the Validation Set are marked in red. In green, the word is shown as it is written in the ground truth transcription.

3. *Results*

The project initially iterated through six models. Fig. 6 shows the size of the training data as number of lines, beside the CERs of each model. Although the results were promising, the CER of the Validation Set remained above 5%, and so work continued on improving the model’s accuracy. This included reinspecting the segmentation of images and proofreading transcriptions (see Griffiths 2022a and Griffiths 2022b).

In October 2022, our HTR model was trained on 269 folios (2310 lines), with validation performed on 27 folios. Using 250 epochs, the trained model

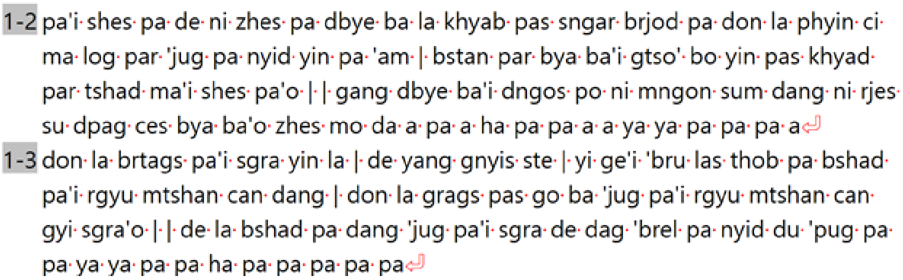


Name: TibSchol 6
Language: Tibetan
Description: Model created by Tibschol, a project that has received funding
Parameters: Nr. of Epochs 250 Omitted Tags unclear
Document Type: Handwrit
Show advanced parameters...
Nr. of Words: 158366
Nr. of Lines: 2194
Save
Show Train Set
ow Validation
how Character
CER on Train Set: 1.86%
CER on Validation Set: 6.48%

Fig. 6 – Results of HTR Models 1-6 trained in Transkribus.

had a CER of 1.15% for the Training Set and a 2.33% for the Validation Set. Satisfied with the results, we then tested the model on other manuscripts written in *drutsa* that were not part of our dataset and were pleased to see that it produced accurate results.

Unfortunately, in November 2022, HTR+ recognition within Transkribus was deactivated. All models trained using HTR+ can no longer be used, although they can be retrained with the PyLaia engine. Training with PyLaia (using the same data set and number of epochs) initially produced poorer results; a CER of 10.20% for the Training Set and 8.80% for the Validation Set. A recurring issue in the generated transcript was a string of random letters that appeared mostly at the end of each line, although sometimes at the beginning (Fig. 7), limiting the use of the model.



1-2 pa'i· shes· pa· de· ni· zhes· pa· dbye· ba· la· khyab· pas· sngar· brjod· pa· don· la· phyin· ci· ma· log· par· 'jug· pa· nyid· yin· pa· 'am· | bstan· par· bya· ba'i· gtso'· bo· yin· pas· khyad· par· tshad· ma'i· shes· pa'o· | | gang· dbye· ba'i· dngos· po· ni· mngon· sum· dang· ni· rjes· su· dpag· ces· bya· ba'o· zhes· mo· da· a· pa· a· ha· pa· pa· a· a· ya· ya· pa· pa· pa· a↵

1-3 don· la· brtags· pa'i· sgra· yin· la· |· de· yang· gnyis· ste· |· yi· ge'i· 'bru· las· thob· pa· bshad· pa'i· rgyu· mtshan· can· dang· | don· la· grags· pas· go· ba· 'jug· pa'i· rgyu· mtshan· can· gyi· sgra'o· |· |· de· la· bshad· pa· dang· 'jug· pa'i· sgra· de· dag· 'brel· pa· nyid· du· 'pug· pa· pa· ya· ya· pa· pa· ha· pa· pa· pa· pa· pa↵

Fig. 7 – Example of errors in transcript produced by PyLaia model before dewarping.

The random string of letters was linked to PyLaia's difficulty in reading curved lines. Fortunately, Transkribus has developed a new tool that tackles this issue: line dewarping (accessed via 'PyLaia advanced parameters'). We found that selecting 'dewarp' significantly improved training results. Our most recent model was trained with a CER of 2.2% on the Training Set and 1.40% on the Validation Set.

4. *Next Steps*

Having successfully trained a model that gives us a CER <3%, it is clear that our research so far indicates that Transkribus works extremely well at recognising historical documents and can be applied to the yet unexplored Tibetan texts in *drutsa* in our corpus.

There are however, two areas of focus to improve our current and future models, as identified through our current validation. The first of these is the occurrence of rarer textual elements, such as numerals, Tibetan rendering of Sanskrit words, and certain symbols and punctuation. Due to their infrequency, the model struggles to recognise these and, instead, produces something that appears similar e.g., a *shad* ། instead of the number one ༡. We have been able to significantly improve the model's recognition of numbers through identifying texts with a higher frequency of numerals and adding these to the training model. The second emerging issue is that the model also transcribes similar looking letters such as *pa* and *ba*, *zha* and *na* with lower confidence, especially with lower quality images and/or damaged or soiled manuscripts. To some degree, these issues will lessen as the model is trained on a greater volume of texts, however, further investigation is required to allow tailored solutions.

The next stage of the project will be to train HTR models for the other scripts found within our corpus. Currently, when tested, the *drutsa* model has a higher CER (>10%) when applied to other scripts. However, used as a base model, it will require significantly fewer pages to train models for other scripts. We estimate around 30 to 50 folios of new ground truth, as opposed to the 300 folios required for the base model. Based on our results with *drutsa*, we are satisfied that our approach of developing multiple HTR models is appropriate for obtaining a reasonably accurate transcription of a large quantity of data in multiple scripts. We would then be able to explore the possibility of training scripts together to create a generic model for Tibetan manuscripts in cursive script.

The development of HTR models that can transcribe multiple Tibetan scripts would open up the possibility of unlocking a vast trove of fully searchable texts that could be available to a much broader audience. To this end, one of our project's aims is to publish our models on Transkribus and make our ground truth datasets publicly available.

Bibliography

Digitizing Jewish Studies (DiJeSt). 2020. Digitize your texts with Transkribus. (last accessed: 30.01.2023). http://dijest.net/digitize-your-texts-with-transkribus/.

dPal brtsegs bod yig dpe rnying zhib 'jug khang (ed.). 2006-2015. *bKa' gdams gsung 'bum phyogs bsgrigs thengs dang po/gnyis pa/gsum pa/bzhi pa*, vol. 1–120. Si khron mi rigs dpe skrun khang.

Griffiths, R. 2022a. Transkribus in Practice: Abbreviations. *The Digital Orientalist* https://digitalorientalist.com/2022/11/01/transkribus-in-practice-abbreviations/. (last accessed: 30.01.2023).

Griffiths, R. 2022b. Transkribus in Practice: Improving CER. *The Digital Orientalist* https://digitalorientalist.com/2022/10/25/transkribus-in-practice-improving-cer/. (last accessed: 30.01.2023).

Kahle, Philip, Sebastian Colutto, Günter Hackl & Günter Mühlberger. 2017. Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, vol. 04, 19–24. doi:10.1109/ICDAR.2017.307.

Merkel-Hilf, N. 2022. Ground Truth data for printed Devanagari [Dataset]. doi:10.11588/data/EGOKEI.

Nockels, J et al. 2022. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science* 22. 367–392. doi:10.1007/s10502-022-09397-0.

O'Neill, A.J. & N. Hill. 2022. Text Recognition for Nepalese Manuscripts in Pracalit Script. *Journal of Open Humanities Data* 8(26). doi:10.5334/johd.90.

Transkribus. 2022. How to Train Baseline Models in Transkribus. (last accessed: 30.01.2023). https://readcoop.eu/transkribus/howto/how-to-train-baseline-models-in-transkribus/.

Wylie, T. 1959. A Standard System of Tibetan Transcription. *Harvard Journal of Asiatic Studies* 22. 261–267. doi:10.2307/2718544.