TibSchol HTR tools

Project Acronym	TibSchol
Project Number	101001002
Authors	Pascale Hugon, Rachael Griffiths
Document version	2.1
Date of version 1	12.10.2023
Date of last update	02.01.2024
List of changes	Update link to GT, add model IDs, add DOIs
License	CC BY-NC 4.0
DOI	10.5281/zenodo.10451396

1. Summary

In the framework of the project TibSchol – "The Dawn of Tibetan Buddhist Scholasticism (11th-13th c.)" (https://www.oeaw.ac.at/projects/tibschol)[1] – hosted at the Institute for the Cultural and Intellectual History of Asia of the Austrian Academy of Sciences, a baseline model for layout analysis (LA) and two handwritten text recognition (HTR) models, for *dpe tshugs* and *'bru tsha* Tibetan cursive scripts respectively, have been trained by Rachael Griffiths on the platform Transkribus (https://readcoop.eu/transkribus/) to produce machine-readable e-texts from handwritten documents. In addition, Python scripts have been created by Rachael Griffiths to pre-process images in order to improve the HTR results. This document presents these tools and explains how the LA and HTR models, which have been made public on the Transkribus platform, can be used by anyone.

2. Pre-processing scripts

a) PDF to JPG converter

This script converts a PDF file into a series of JPG images. It utilizes pdf2image library to perform the conversion and PyPDF2 for splitting PDFs for processing, if needed.
https://github.com/ERC-TibSchol/pdf-to-jpg
DOI: 10.5281/zenodo.10450672.

b) Image processing script

This code is designed to enhance image quality by improving resolution, denoising, and sharpening. It requires the OpenCV library to perform image processing tasks.
https://github.com/ERC-TibSchol/image-processing
DOI: 10.5281/zenodo.10450684.

[1] The project TibSchol has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101001002). These tools are being shared within the project team's responsibility. TibSchol must not be held responsible for the accuracy of the output when using these tools or for future changes that would have an impact on their operation. The European Research Council or the European Commission must not be held responsible for their further use.

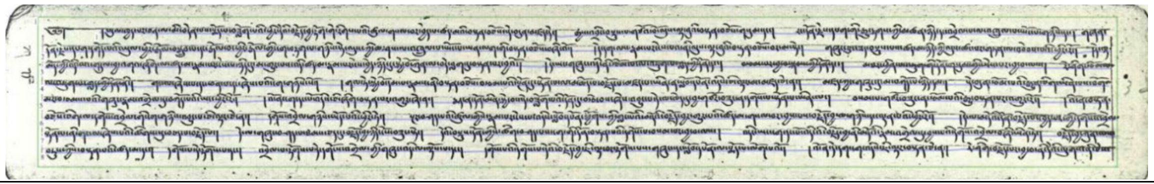
c) Image segmentation scripts (segment in four and segment in two)

These codes allow one to cut an input image into two or four equal horizontal segments and save them as separate images. They utilize the Python Imaging Library (PIL) to perform the image segmentation.
https://github.com/ERC-TibSchol/image-segmentation
DOI: 10.5281/zenodo.10450698.

3. TibSchol models

a) Layout analysis model Tibetan Pecha (ID:54306) – CER 6.64%

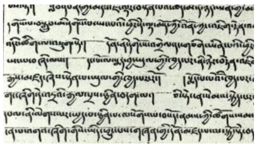
This baseline model has been trained on 1699 folios from sources being explored in the project. It was trained to recognize horizontal baselines only (i.e., titles, main text and glosses).



Example of layout recognition for fol. 2a from gTsang nag pa's *Tshad ma rnam par nges pa'i ti ka legs bshad bsdus pa*, preserved at Otani University. Image: http://purl.bdrc.io/resource/W1KG12371.

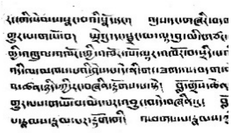
b) HTR model Tibetan cursive (Drutsa) (ID:54525) – CER 1.40%

This model is tailored towards transcribing the handwritten Tibetan cursive script known as Drutsa (*'bru tsha*). Ground truth data consists of 466 folios from a selection of 19 Tibetan treatises being explored in the TibSchol project. 422 folios were used in the training set and 44 in the validation set. Transcripts use the extended Wylie transliteration system in Roman alphabet. Abbreviations were transcribed as they appear in the manuscripts, for a list of abbreviations see https://github.com/ERC-TibSchol/abbreviations; DOI: 10.5281/zenodo.10450653.



c) HTR model Tibetan cursive (Betsug) (ID:54935) – CER 3.6%

This model is tailored towards transcribing the handwritten Tibetan cursive script known as Betsug (*dpe tshugs*). This model uses the Drutsa model ('Tibetan cursive (Drutsa)') as a base model. Ground truth data consists of 93 folios from a selection of 8 Tibetan treatises being explored in the TibSchol project. 85 folios were used in the training set and 8 in the validation set. Transcripts use the extended Wylie transliteration system in Roman alphabet. Abbreviations were transcribed as they appear in the manuscripts, for a list of abbreviations see https://github.com/ERC-TibSchol/abbreviations. The GT also includes 360 images of abbreviations kindly provided by MonlamAI (https://monlam.ai/).



More information on the training of these models:

- Griffiths, Rachael, Transkribus in Practice: Improving CER. *The Digital Orientalist*, 25 October 2022. (https://digitalorientalist.com/2022/10/25/transkribus-in-practice-improving-cer/)
- Griffiths, Rachael, Transkribus in Practice: Abbreviations. *The Digital Orientalist,* 1 November 2022. (https://digitalorientalist.com/2022/11/01/transkribus-in-practice-abbreviations/)

4. How to use the models

The models are part of the public models on the platform Transkribus (https://readcoop.eu/transkribus/). For detailed instructions and answers to all questions related to the use of the Transkribus platform, in particular creating an account, uploading documents, running recognition, and exporting results, please refer to the **READ Co-op/Transkribus "Helpful resources"** (Video tutorials, How-to Guides, FAQ…). Credits are only needed for text recognition proper.

The LA model "Tibetan Pecha" (ID:54306) can be selected from the list of public models for "Layout recognition." The HTR models "Tibetan cursive Drutsa" (ID:54525) or "Tibetan cursive Betsug" (ID:54935) can be selected from the list of public models for "Text recognition."

- To improve recognition performance, we recommend transforming PDF documents into images (.jpg) and pre-processing the images before uploading them. The tools described in §2 above can help you achieve this.
- For better results, it is recommended to apply the specific LA model "Tibetan Pecha" before running the recognition with one of the two Tibetan HTR models (note: users need to sign into the Transkribus App to be able to select a specific LA model).
- Because running text recognition consumes credits, we advise testing the results on a few folios before processing large documents.
- On the platform, you can make corrections to the layout before running text recognition, and make corrections to the transcription that has been generated before exporting, but this will not improve the respective models.
- The HTR were trained on texts transcribed with the Wylie transliteration system and use this system for the output of text recognition. If you prefer working with Tibetan Unicode, there are several free online converter tools and software that allow you to transform Wylie in Tibetan Unicode characters. (See for example: https://www.thlib.org/reference/transliteration/wyconverter.php, https://github.com/karmapa > http://karmapa.github.io/wylie/demo/index.html)

5. How to make your own model

If our Drutsa and the Betsug models do not generate satisfactory results with the documents you want to process, or if you are aiming at a higher accuracy, you can train your own model from scratch, or by building on the TibSchol models. The Ground Truth (data set and metadata) is accessible on: https://github.com/ERC-TibSchol/Tibetan-Cursive-GT; DOI: 10.5281/zenodo.10450635. In that case, you are kindly requested to mention the TibSchol base model in your own model description.