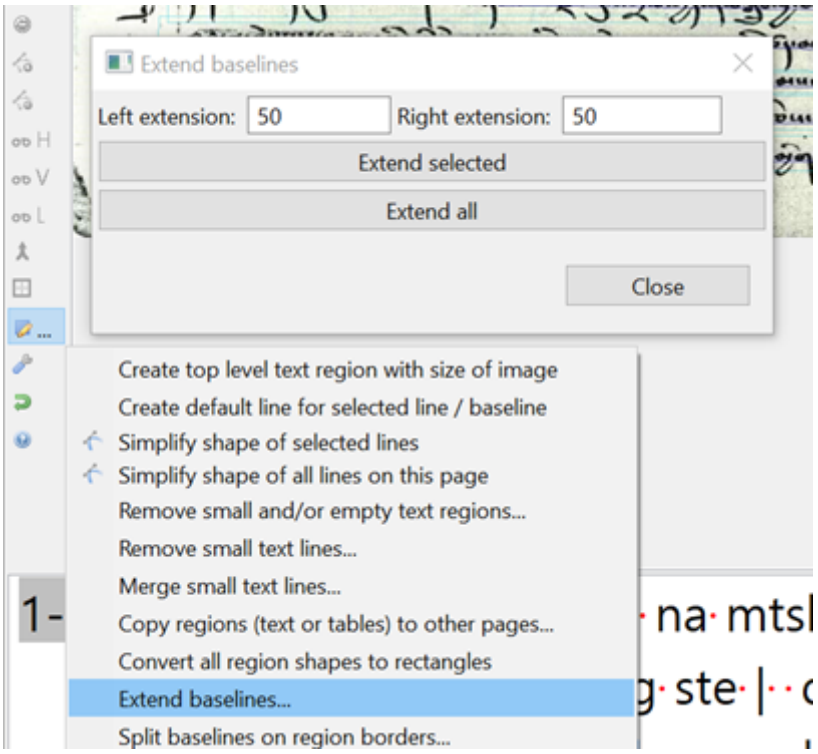
Transkribus in Practice: Improving CER

Since joining the ERC-funded project, ***The Dawn of Tibetan Buddhist Scholasticism (11th-13th C.)*** **(https://www.oeaw.ac.at/ikga/tibschol)**(TibSchol),[1] at the Austrian Academy of Sciences in 2021, I have been experimenting with ***Transkribus*** **(https://readcoop.eu/de/transkribus/)**. The goal is to see if *Transkribus* can train Handwritten Text Recognition (HTR) model(s) that can automatically process Tibetan cursive (*dbu med*) manuscripts of works from the 11th to 13th centuries. If successful, this would make a large amount of the early bKa' gdams pa (བཀའ་གདམས་པ་) scholastic corpus text searchable. This article will not be providing instructions on how to train a model with *Transkribus*, as there are **extensive guides (https://readcoop.eu/transkribus/resources/how-to-guides/)** already available. Instead, in this and subsequent posts, I will outline my experiments to improve the accuracy of our model.

Our existing model – with a dataset of 295 pages, representing 2,194 lines – had a Character Error Rate (CER, the percentage of characters that have been transcribed incorrectly by the model) of 1.68% for the Training Set and a 6.48% for the Validation Set (pages not used to train the HTR, which instead evaluate the performance of the model). Although these results are considered very efficient (anything below 10% is), there is still room for improvement. I was curious to know if the CER could be easily reduced.

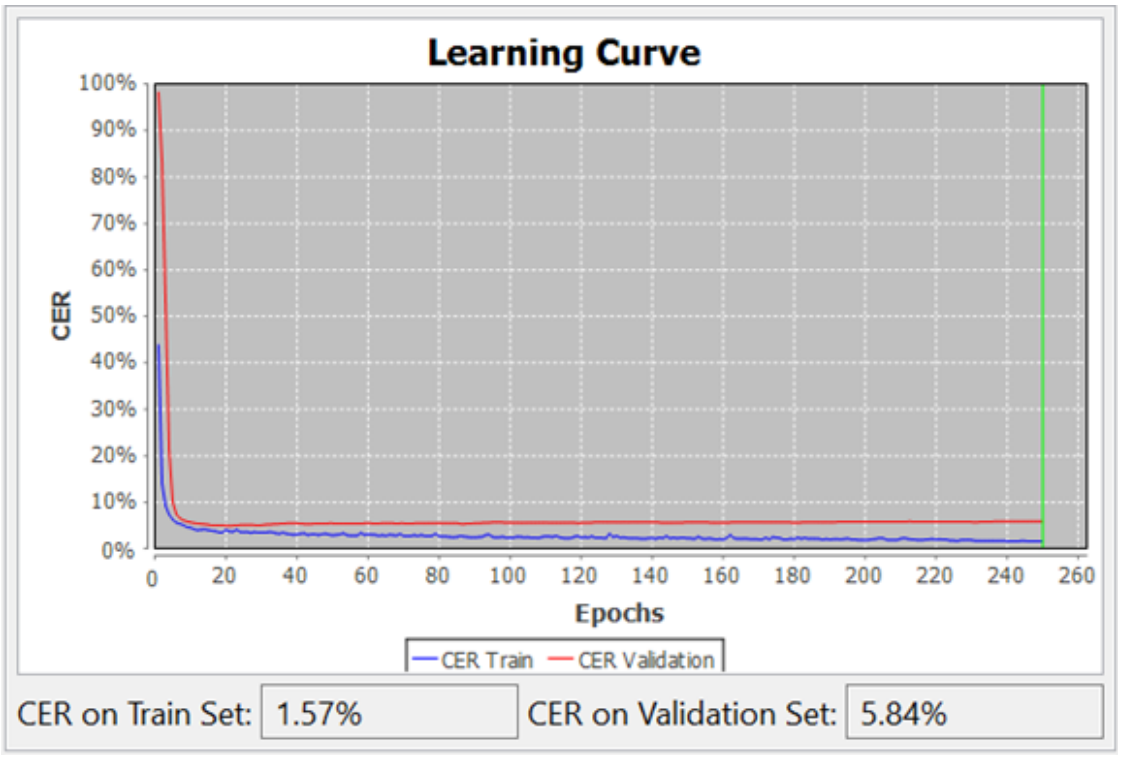
My first step in trying to improve the model's accuracy was going back to inspecting the text regions and baselines that the layout analysis had added to the manuscripts. This was to ensure that all text was included in the text regions, and all baselines extended fully along the handwritten text line (baselines can be horizontal and/or vertical), as fixing errors in these areas could cause significant reductions in CER. There were some instances where the baseline missed the beginning and/or end of the line. Instead of correcting each of these by hand, which is a fiddly and time-consuming process, I used the "extend baselines" tool, which allows you to extend each end of the baseline by a chosen amount. The tool automatically suggests an extension of 50 pixels, which I have found to be too much, as it begins to overlap with marginalia, folio numbers, etc. After some initial testing, I found that an extension of 5 to 10 pixels is ideal for the materials in our project. Unfortunately, this tool only extends the baselines on the selected page, so it still requires some time and effort (but still far quicker than doing it manually).



(https://digitalorientalist.files.wordpress.com/2022/10/screenshot-1.png)

Extend baselines can be found in the "Canvas" menu left of the image

After correcting text regions and baselines, a new HTR model was generated using the same dataset as before. The new model yielded lower CERs in both the Training and Validation Set, a promising result!



(https://digitalorientalist.files.wordpress.com/2022/10/screenshot-2.png)

Learning curve and CER of the TibSchol model after correcting text regions and baselines

The next step was much larger in scope; checking our transcripts against the images to ensure that they were as faithful as possible to the manuscripts. We were fortunate to have transcriptions of some texts already available to us in Tibetan transliteration, which initially sped up the process of creating our training data. However, when I began the process of checking these against the images, I realised that different transcription conventions had been used by different transcribers, for example, some transcribed གཡས (right) as *g.yas*, following the Wylie transliteration method, while others used *g-yas*. Although seemingly small, every transcription error produced by the model is considered a fully-fledged error and is included in the CER. Moreover, some had chosen to write abbreviations in full, while others transcribed what they read on the image, for example, *laso* or *la sogs* (ལ་སོགས; and so forth). I will discuss abbreviations in more detail in a forthcoming post, as it deserves more than a couple of lines. For now, it is clear that these variations within the training data necessitated a review of all the transcriptions in detail. This takes a great deal of time, not only because of the amount of data to check and correct, but also because I chose to document our transcription and abbreviation conventions for reference in future parts of the project in parallel. At the time of writing this article, I have reviewed 177 pages of transcripts (60% of our training data).

Because this process is time-consuming, I wanted to ensure that it was having an impact on our CERs to confirm that this time was being used effectively. I created four different models, each using the same dataset as our existing model. The models were created as follows:

- *Model A* after 40 pages were manually checked
- *Model B* after 80 pages were manually checked
- *Model C* after 120 pages were manually checked
- *Model D* after 160 pages were manually checked

The results are summed up below:

Model name	**No. of pages checked**	**CER% for Training Set**	**CER% for Validation Set**
Model A	40	1.39%	4.28%
Model B	80	1.35%	4.45%
Model C	120	1.18%	4.73%
Model D	160	1.15%	2.33%

Encouragingly, the CER for the Training Set continues to drop and could fall below 1%! However, the CER for the Validation Set is less consistent, having initially dropped by over 1.5% before increasing again and then falling by over 2.5%. The increase could be linked to the standardisation process, where our chosen transcript conventions initially resulted in more characters being transcribed incorrectly by HTR. The expectation is that as more of our transcripts are corrected and standardised, the more accurate our model will perform on our Validation Set. The results of *Model D* are very promising, so I will continue to check the remaining pages of our transcripts. Watch this space!

Based on my experience, I have reached two main conclusions: 1) for those preparing to train a model, if you are using existing transcripts, check them thoroughly before adding them to *Transkribus.* It could save you a lot of time later on in the process. 2) For those looking to improve their CERs, checking the page layout and transcriptions can yield positive results, but it is laborious. A CER of 10% or lower is recommended for efficient automated transcription. If your CER falls within the range, perhaps a more efficient use of time would be correcting the automated transcriptions instead, especially those whose goal is producing a critical digital edition.

Endnotes

[1] 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.

(https://digitalorientalist.files.wordpress.com/2022/10/eu-logos.png)

Published by

Rachael Griffiths

View all posts by Rachael Griffiths

OCTOBER 25, 2022OCTOBER 27, 2022 LEAVE A COMMENT BUDDHIST STUDIES, HTR, TIBETAN STUDIES, TRANSKRIBUS