

From Large and Complex Manuscript Collections to Searchable eTexts: the Case of PaganTibet

Rachael M. Griffiths^{a,b} & Marieke Meelen^b

^aEPHE-PSL, ^bUniversity of Cambridge

Abstract

his article presents a comprehensive workflow for creating searchable eTexts from complex manuscript collections, using the PaganTibet corpus - made up of over 100,000 manuscript pages - as a case study. The corpus presents multiple challenges including its size, diverse handwritten styles (umé), and extensive use of abbreviations. We outline a collaborative approach integrating careful planning, multiple rounds of quality control, AI-enhanced Handwritten Text Recognition (HTR) technology, and iterative error analysis. Comparing HTR model performance on challenging test sets, we demonstrate that models trained on carefully curated, diverse training data significantly outperform those trained on larger but more homogeneous datasets. Moreover, our analysis shows that while conventional evaluation metrics such as Character Error Rates (CERs) are useful benchmarks, they do not always reflect the practical performance of HTR models.¹

1. Introduction

In the last few decades, large collections of Tibetan manuscripts have come to light in remote parts of the Sino-Tibetan borderlands, which Tibetan media in China reported to date from the Tibetan imperial period. The publications of facsimiles based on these manuscript caches

¹ The research for this article is partly funded by the ERC Advanced Grant “Reconstructing the Pagan Religion of Tibet (PaganTibet)”, project 101097364. Views and opinions expressed are those of the authors and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We would like to express our gratitude to the journal editors, the internal PaganTibet project members, our collaborators at the BDRC and Esukhia, and especially our annotators at Triten Norbutse Institute in Kathmandu: Sherab Chokgyal (Team Leader), Tsewang Drukgyal, Tsultrim Gyaltsen (Team Leader), Tsognyi Gyatso, Tritsuk Lhundup, Tsultrim Palsang, Palgyi Wangchuk, Tsukphud Rabsal, Tsugphud Woeser, and Sherap Woser.

found in private houses are now generally referred to as the “Leu collections”² after Ngawang Gyatso’s 2005 article on a local tradition of lay priests called *leu* (*le’u*) in the Minshan mountain range (on the border of Gansu and Sichuan, China).³ As Berounský & Ramble (2026) discuss, these Leu collections present a mixture of texts, some referring to Buddhist and Yungdrung Bön religious practices, myths and rituals. Many, however, appear to describe autochthonous traditions prior to the advent of Buddhism and the varieties of indigenous Bön, which we, following Ramble (1998), refer to as “Pagan”.⁴ Apart from the well-known Dunhuang⁵ and Gathang⁶ texts, there are no written records of pre-Buddhist times. The topics of the rituals and the codicological features in the Leu texts are also found in some of the Dunhuang and Gathang texts, as well as in another recently discovered collection of ritual texts of the Baima who live in Pingwu, Songpan (Tib. *Zung chu*), Jiuzhaigou (Tib. *Gzi rtsa sde dgu*) in Sichuan, and in the counties of Wenxian and Zhouqu (Tib. *’Brug chu*) in Gansu, China.⁷ In addition to this, one of the “portals”⁸ of Yungdrung Bön texts is a repository of similar Pagan myths and rituals known as Black Water (*chab nag*), which are also found in compendia of ritual texts used by Bönpo village lamas in Mustang and Dolpo (Nepal).

The newly discovered Leu collections, containing at least 35,000 folios, combined with the relevant Baima, Dunhuang, Gathang and Black Water texts, comprise a growing corpus that to date contains >100,000 pages. These collections, which we will henceforth refer to as the “PaganTibet corpus” offer great potential for historical, anthropological, religious as well as philological and linguistic research for the time period. Berounský & Ramble (2026), however, list a number of reasons why researchers have not tapped into this rich resource yet, including the lack of searchable transcriptions (i.e. eTexts), the idiosyncratic and wide variety of forms of the *umé* (*dbu med*) scripts in which they are written, the highly irregular orthography and abundant use of contracted forms as well as unusual symbols, together with the lack of dictionaries and other resources for the specialist terms and unfamiliar narrative types and structures.

² The published facsimiles are generally referred to as “Leu 2011-30”, “Leu 2013-60”, “Leu 2016-10”, and “Leu 2024-100” following the year of publication and the number of volumes in the collections. Further publications are currently still in preparation.

³ See Ngawang Gyatso (2005).

⁴ Ramble (1998:124) provides further justification for the use of the term “Pagan”.

⁵ See <https://idp.bl.uk/about/projects/tibetan-manuscripts/>.

⁶ See *dGa’ thang ’bum pa nas bon gyi gna’ dpe* (2007).

⁷ See La Xian (2022).

⁸ For an overview of the eclectic nature of the Bön religions, see, among others, Snellgrove (1967:13); Karmay (1998 [1975]:119) and Kværne (1995:19).

In this article, we first introduce a comprehensive, AI-enhanced workflow for creating searchable eTexts for challenging collections like the written⁹ PaganTibet corpus that can form a blueprint for other large manuscript projects (Section 2.1). We then discuss specific issues of data collection and curation (Section 2.2) as well as codicological and palaeographical challenges (Section 2.3). Following this, we present the results of our Handwritten Text Recognition (HTR) models in Section 3, with an in-depth error analysis of a selection of Leu texts comparing the PaganTibet models to other available options. In Section 4 we summarise our findings with lessons learnt and concrete suggestions for future research.

2. Challenges for Large and Complex Text Collections

In this section we address the major challenges that large amounts of difficult text collections present, drawing on our experience of working with the PaganTibet corpus. We first deal with the size, suggesting a collaborative workflow, before diving into specific issues with data collection and curation as well as palaeographical and codicological challenges.

2.1 Issues with Large Collections

The most obvious challenge of dealing with large amounts of manuscripts lies in the practical difficulties of transcribing the texts, either directly or from images. Even highly skilled and experienced philologists need time to manually transcribe dozens, if not hundreds or thousands of folios, which is impossible, especially within a time-sensitive project. Meelen et al. (2024) explain why well-known automated solutions, in particular Optical Character Recognition (OCR), are not suitable for Tibetan umé due to the handwritten text, which presents ligatures as well as connecting and overlapping characters (unlike, for example, typed Latin or Tibetan Unicode scripts, that feature white space between characters). Instead, Griffiths (2024) presents a method for digitising Tibetan umé line by line using an HTR model.¹⁰ HTR models are deep neural networks (i.e. “AI systems”) that learn how to transcribe handwritten text by training on examples of images paired with

⁹ In addition to written texts, within the PaganTibet project we also deal with audio-visual data in the form of ritual videos and interviews with Leu and Baima priests. These audio-visual collections are part of the broader PaganTibet corpus, but this article only deals with the written component. For workflows and tools on how to transcribe audio-visual data, see Chirkova et al. (2025).

¹⁰ HTR models can also be trained to automatically transcribe handwritten uchen (*dbu can*), see Werner & Viehbeck (2025), and printed uchen, see Erhard (2025).

their correct transcriptions called “Ground Truth”. Through this training, HTR models learn statistical patterns of computer vision (i.e. dark/light/coloured pixels) that enable them to map visual features (i.e. handwritten letters/words/lines) to corresponding text sequences. Once the models are trained, they can process new ‘unseen’ batches of images at scale to reduce the time spent on creating transcriptions. HTR models only work if the text regions and baselines can be identified automatically as well. In order to recognise where the relevant material is, so-called “Layout Recognition” or “Layout Analysis” models have to be applied or trained first so that they can inform the HTR model where to ‘look’ for text to transcribe (see Fig. 1).

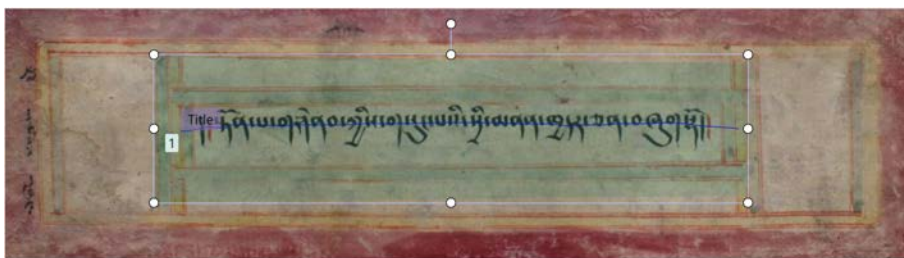


Fig. 1 — Transkribus model prediction with a text region (in green) and baselines (in blue). Folio taken from <https://kalpa-bon.com/texts/sri/sri-mnan-khyad-par-can>.

However, large manuscript collections - especially disparate collections - present challenges that extend beyond the technical aspects of utilising both HTR and Layout Analysis models. The scale of the PaganTibet corpus creates several interconnected issues.

First, producing sufficient Ground Truth (i.e. accurate transcriptions used to train HTR models, also called “Gold Standard” or simply referred to as “training data”) becomes a major bottleneck. Manual transcription and the necessary checking of even a small percentage of the PaganTibet corpus requires substantial resources and time. This is particularly challenging when the texts demand specialist expertise, where the pool of capable transcribers is limited. Moreover, ensuring the manuscripts used for Ground Truth best captures the diversity of scripts and breadth of complex features within the corpus presents an additional challenge. Second, maintaining consistency across such a large corpus is difficult. Due to the aims of the project, the eTexts produced by HTR will be diplomatic, i.e. as faithful as possible to the manuscripts, to facilitate in-depth philological, linguistics and comparative

research.¹¹ As such, the workflow necessitates the ongoing identification and documentation of complex cases, such as abbreviated forms and instances affected by Unicode limitations (discussed in Section 2.3), and the development of consistent conventions for their representation. Third, the PaganTibet corpus exhibits substantial internal variation. Different script hands, states of preservation, and image quality mean that an HTR model may perform well on some materials while failing on others. Consistent monitoring of model performance through qualitative analysis is therefore essential for ensuring the model is as robust as possible.

To address these challenges, we implemented a collaborative workflow. This workflow integrates three key partners: the internal PaganTibet team, partner organisations specialising in digital preservation and language technology (BDRC and Esukhia), and specialist annotators at the Triten Norbutse Institute (TNI), Kathmandu. The internal team handles image selection, model training, and final quality control. Partner organisations provide platform development, data management, and open access repository services. The annotation team is divided into two groups: one conducting manual transcription to create initial Ground Truth, and another correcting automated HTR output.¹²

The workflow uses a tiered quality control system with three levels of review. This proved essential, with the final review stage identifying as many as 10-15% of errors remaining. However, even after this triple-check process, small errors were found in the training data. During model training, Transkribus displays the model's predictions alongside the Ground Truth, and in some cases the model's output was correct while the Ground Truth still contained errors (see Fig. 2). Locating and correcting these errors meant that approximately 10% of the training data underwent a fourth review and correction round. This highlights the importance of continuously checking training data quality, as even rigorous multi-stage review processes cannot catch all errors. We will return to this issue when discussing the results in Section 3 below.

¹¹ HTR models can, to an extent, be trained to produce normalised eTexts, but it depends on the extent of abbreviations, spelling variations etc. and the amount of Ground Truth available - see the Transkribus website for more on this: <https://help.transkribus.org/data-preparation>.

Normalised versions of the transcriptions, which standardise spelling variations, expand abbreviations etc. to create consistent, searchable texts, are currently in development. For more information, see Griffiths & Meelen (2025b) and Meelen & Griffiths (2026).

¹² For a detailed discussion of workflow implementation, training procedures, and platform usability testing, see Meelen & Griffiths (2025b).

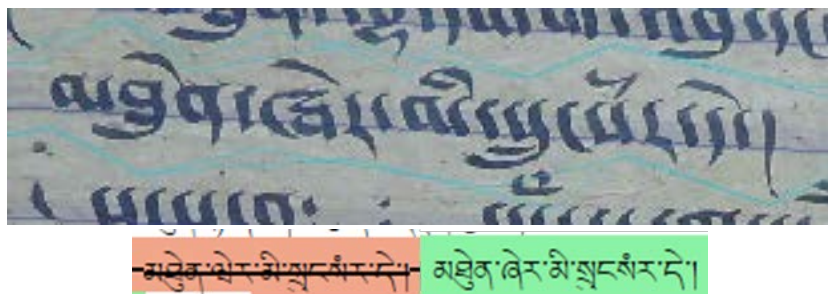


Fig. 2 — The ‘incorrect’ model prediction (in red) alongside the ‘correct’ Ground Truth (in green). The model has accurately transcribed ཞེ (wer), which appears as ཞེ (zher) in the Ground Truth.

Beyond quality control, the workflow's efficiency depended on appropriate technological infrastructure. The use of PechaTools,¹³ a platform developed by Esukhia specifically for Tibetan transcription and related tasks in digital humanities, facilitated coordination through integrated features including task assignment, error highlighting (allowing reviewers to mark mistakes in red for annotators to learn from), and a built-in dictionary of diplomatic rendering conventions. Regular updates to transcription manuals and training materials¹⁴ addressed recurring issues with complex cases.

To test the robustness of our workflow and training materials, we conducted an inter-annotator agreement (IAA) pilot analysis on a small sample of 185 lines drawn randomly from different texts across the corpus to compare how annotators transcribed these lines. Results showed strong agreement overall, and will be discussed in more detail in Section 3 below. The IAA and review stages were also useful for identifying recurring transcription errors and inconsistencies, for example the use of the *anunāsika* ས (U+0F84), which is not part of our transcription standards, instead of the *anusvāra* ས (U+0F7E). To address these issues, we developed a postprocessing script that applies rule-based corrections to our transcriptions.¹⁵ Corrected transcriptions were then incorporated back into the Ground Truth before model retraining, creating an iterative feedback loop that progressively reduces errors and improves future models.

Ground Truth data (triple-checked transcriptions) was exported from PechaTools and uploaded to Transkribus, a web-based platform for training and applying HTR models, for model training.¹⁶ We chose Transkribus for several reasons: it is user-friendly with a web-based

¹³ Also known as ‘BuddhistAItools’, see <http://buddhistai.tools/>.

¹⁴ See the HTR manual (Meelen & Griffiths 2025a) and the HTR Cheatsheet (Griffiths & Meelen 2025a).

¹⁵ See <https://github.com/pagantibet/htr>.

¹⁶ See <https://app.transkribus.org/>.

interface that requires no technical setup or server access, and it already hosts publicly available Tibetan umé models that provided a useful starting point for our work.¹⁷ However, there are limitations: models trained on Transkribus are publicly available but can only be used within Transkribus itself. It is also a paid platform. While 50 free credits are provided monthly (sufficient to process approximately 40 images), processing larger batches requires paid subscriptions.

2.2 *Issues with Data Collection and Curation*

2.2.1 *Issues collating existing collections*

Corpora are often created from a range of different existing collections of images. Both private collections and published facsimiles can pose some problems when creating a coherent corpus. As Berounský & Ramble (2026) point out, some of the published facsimiles of the Leu collections contain images of texts that are printed upside down or in the wrong order, i.e. starting with the end of the text working back to the title page later on. In other cases, longer texts are interspersed with single folios of other texts. This issue arises for the Leu 2011-30 collection in particular. Although the later Leu 2013-60 and Leu 2016-10 collections are well-organised, there are no content catalogues for any of them.¹⁸

When it comes to private collections, a major issue with uncatalogued and undocumented sets of images is the risk of duplicates. Digital cameras often have similar file-naming conventions and when SD cards are changed and/or images are copied to multiple devices, identical images are often stored multiple times in different folders and/or photographs of different manuscript pages end up with identical names. When the file names are identical, these can be easily filtered by searching for and deleting duplicates, as well as renaming those that depict/contain different content. Sometimes, however, file names differ but the image is exactly the same. In those cases identical images can only be identified through software that compares and flags duplicate images. The most challenging cases arise when photographers have taken multiple pictures of the same page, for example because their first attempt was out of focus. If the lower-quality images are not immediately deleted and end up in the final collections, they can only

¹⁷ The first publicly available Tibetan models on Transkribus were trained by the TibSchol project at the Austrian Academy of Sciences: *The Dawn of Tibetan Scholarship* (11th–13thc.) (H2020-Cog-101001002), <https://cordis.europa.eu/project/id/101001002/reporting>. For more on these models, see Griffiths (2024).

¹⁸ See Griffiths & Meelen (2025b) for a workflow on creating content catalogues semi-automatically, including a pilot case study on identifying Pagan texts.

be found when going through the entire collection, image by image. This works for small collections, but is much harder when thousands of images need to be checked.

2.2.2 Image Quality and Technical Considerations

The quality and technical specifications of manuscript images directly impact HTR performance. High-quality images not only produce better recognition results but also eliminate the need for preprocessing stages.¹⁹ While collating batches of images for transcription, we encountered several challenges that drew attention to the importance of careful digitisation.

Image quality issues such as poor focus or inconsistent lighting substantially degrade HTR performance. Blurred text makes it difficult for models to distinguish individual characters, particularly when dealing with handwritten scripts. While some preprocessing techniques can improve contrast or sharpness, these cannot fully compensate for poor quality at capture. Similarly, when manuscript pages are photographed stacked on top of each other, underlying pages may be visible. This creates visual interference that can confuse Layout Analysis (i.e. marking up where text appears on an image; see Fig. 3) and HTR models, which may attempt to recognise text from other layers.²⁰ Images taken at angles rather than directly overhead can also distort text, particularly near page edges. Although Transkribus has optional advanced parameters, such as binarisation (turning an image into black and white to help the AI focus on key aspects of the image) and dewarping (correcting distortions to produce a flatter image), which cut out some preprocessing, these settings cannot fully remedy quality issues.

¹⁹ Preprocessing enhances images to allow further processing - this can include deskewing, reducing image noise, and increasing image size and resolution. One example of a preprocessing script developed for Tibetan language texts is Divergent Discourses Custom Image Preprocessing (Sabbagh et al. 2004), developed by the *Divergent Discourses* project <https://research.uni-leipzig.de/diverge/>. For more on the background and development of this preprocessing script, see Sabbagh (2025).

²⁰ A custom layout analysis model was trained on Transkribus to navigate issues such as stacked manuscript pages. The model was trained on 15,006 images (130,869 lines). It is publicly available: 'Tibetan Manuscript 3' (Transkribus model ID: 229409).

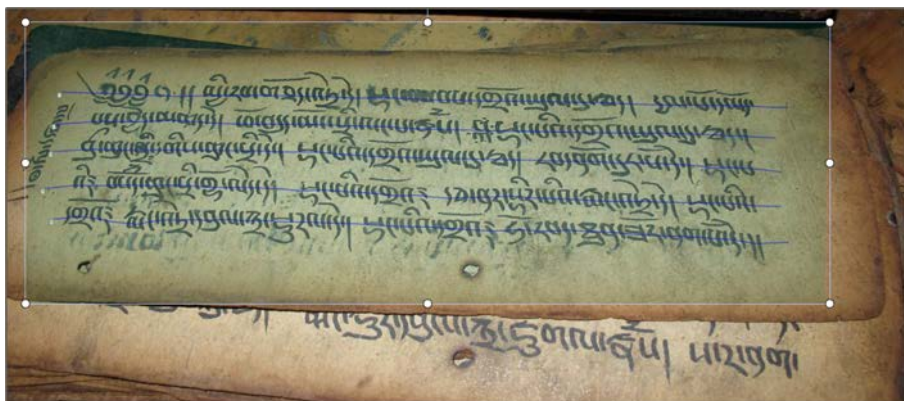


Fig. 3 — An example of stacked manuscript pages. A custom layout analysis model was developed which learnt to ignore the underlying pages, marking only the text region (large green box) and baselines (blue lines) of the top page. © Ngawang Gyatso.

Image metadata presents another technical consideration. Digital images contain EXIF metadata that includes orientation information. This created an unexpected challenge when images photographed at a 90-degree angle were manually rotated, but the underlying EXIF orientation data remained unchanged. When these images were uploaded to PechaTools, the platform automatically read the metadata and displayed them at the incorrect angle. Images could not be rotated within the PechaTools interface. Resolving this required developing a Python script to strip the EXIF orientation data from all affected images.²¹ While implementing this solution was straightforward, identifying the problem and processing thousands of images added considerable time to the workflow. Clear digitisation protocols were subsequently established for data collectors and curators to ensure that new images captured met consistent image quality standards to reduce downstream processing time and improve final HTR accuracy.

The composition of manuscript photographs introduces an additional set of considerations. In our case, many images contain two folios per photograph, which reduces processing costs given that platforms such as Transkribus typically charge per image. This approach, however, necessitates an extra workflow stage in which images must be cropped before or after HTR processing. More broadly, our experience demonstrates that decisions made at the point of capture, such as camera angle, image composition, and metadata handling, have cumulative effects across the entire HTR pipeline. For projects working with large or heterogeneous manuscript collections, the establishment of clear digitisation standards and guidelines is essential not only for

²¹ See <https://github.com/pagantibet/htr>.

optimising final HTR accuracy but also for ensuring workflow efficiency and long-term sustainability.

2.3 Palaeographical and Codicological Challenges

Alongside the technical and logistical challenges discussed above, the PaganTibet corpus presents substantial palaeographical and codicological challenges. For example, the corpus encompasses a wide range of umé, reflecting different styles, time periods, and scribal hands. We find considerable variation in grapheme formation, proportions, and spacing - even within a single page, where two or more scripts are used (see Fig. 4).

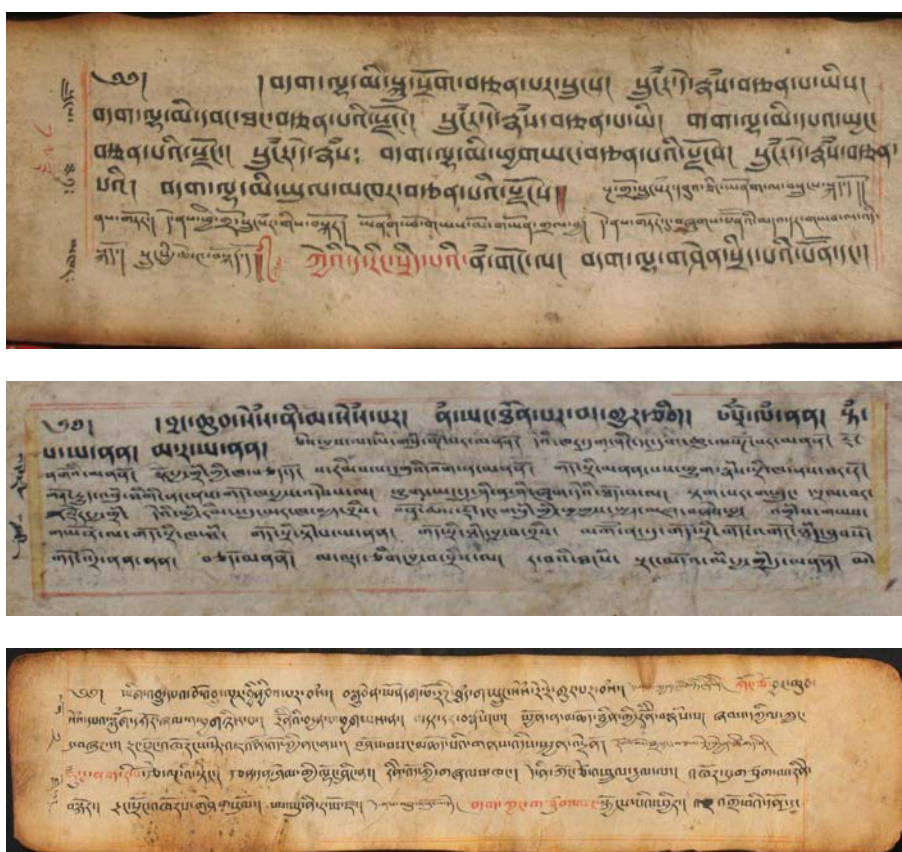


Fig. 4 — Examples of folios with multiple scripts used. Taken from <https://kalpa-bon.com>.²²

²² From top to bottom: image 00205_0011r in <https://kalpa-bon.com/texts/leu/bya-rdang-srung-mkhar>; image 00071_0004r in <https://kalpa-bon.com>

This diversity is further illustrated in Fig. 5 through examples of ཤེང་ (*sheng*) from different manuscripts in the corpus. Although the variations may appear small, they pose a significant challenge for HTR models, which have to learn to recognise the same grapheme across multiple realisations. Insufficient representation of this diversity in training data leads to reduced accuracy on manuscripts whose script differs from the dominant forms within the Ground Truth. For this reason, it can be beneficial to train separate HTR models on each hand or script. This approach was taken by the TibSchol project, which produced dedicated models for Drutsa (Transkribus model ID: 54525) and Betsug (Transkribus model ID: 54935).

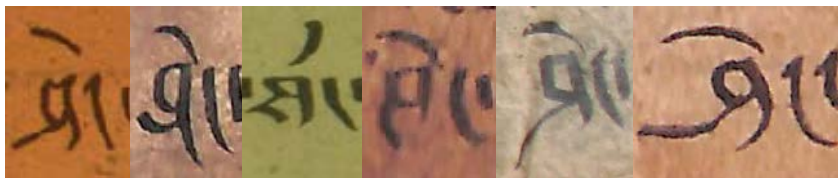


Fig. 5 — Examples of ཤེང་ (*sheng*) found in manuscripts within the PaganTibet corpus.

Given the scale of the project and the fact that we continue to collect new materials, the full extent of script variation remains unknown. We therefore opted to train a single generalised model capable of handling the full breadth of umé varieties. This approach provides greater flexibility for processing new manuscripts, but places considerably higher demands on the Ground Truth. The general model needs to be trained on significantly more data and exposed to a wide and balanced sample of styles to ensure reliable performance across our diverse manuscripts.

This palaeographical variability is further compounded by the extensive use of abbreviations, which constitute a prominent and challenging feature of our manuscripts. Types of abbreviation vary from simple consonant stacking, such as གཉིས་ -> གཉིས་སྟེ (gnyiste, gnyis ste), to more complex forms, such as ཐྱིང་ -> ཐུགས་དམ་བཟང་ (th+yaM dang, thugs dam bskang) and ཐུན་ལྷ་འདྲ་དཀར་ (sh+yen+hador, gshen lha 'od dkar), all of which need to be captured as precisely as possible in our diplomatic transcriptions. The PechaTools platform supported this process through an integrated dictionary of rendering conventions, although

not all abbreviations were initially listed.²³ The frequency of abbreviations naturally varies, with some appearing repeatedly throughout the corpus, such as ཐམ (thaMd, thams cad), རིམ (rine, rin chen), and རེམ (rdo+e, rdo rje), while others occur only occasionally, such as ཇམ (c+tir, ci ltar), གཤེར་མེས་ (gsereMs, gser skyems), and རྒྱུ་ལོ་འདྲེ་ལོ་འདྲེ་ (rgyu+o+on+yi, rgyu 'byor yon gyi). This uneven distribution means that common abbreviations are well-represented in our Ground Truth while infrequent abbreviations remain challenging for models to recognise, especially as there are fewer examples of these abbreviations written in different umé styles (see Fig. 6, for example), posing additional recognition challenges.



Fig. 6 — Examples of ཇམ (c+tir), the abbreviated form of ཇམ་ལྷ་མོ་ (ci ltar), and how it varies across manuscripts.

We also find a range of symbols, some decorative and others used to represent particular words or concepts (see Fig. 7), which present several difficulties. Most have no Unicode equivalent, requiring conventions that approximate the symbol's visual form while remaining meaning-neutral. These conventions must also be simple to reproduce and to find on a Tibetan keyboard for consistent representation across transcriptions. Moreover, like abbreviations, their frequency varies, with some appearing more regularly than others. We continue to encounter new symbols as we process more manuscripts, making it difficult to anticipate the full range of forms that HTR models should be able to recognise. To address these issues, we introduced the symbol × (U+0F1D)²⁴ in our transcription conventions to mark decorative elements and instances where graphemes or symbols are unclear and/or have no current transcription convention. Although useful for noting sticking points that we can easily find and address (see below for a discussion on tags), this does not fully solve the issue. When an HTR model encounters unknown symbols in new manuscripts, it will attempt to transcribe them based on visually-similar patterns from the

²³ We will extend this abbreviation dictionary as part of the next stage of the project's workflow. For more see Griffiths & Meelen (2025b).

²⁴ This symbol is easily accessible to annotators as it is found on the Tibetan Monlam keyboard (Shift+6). This keyboard is free to download and can be installed from: <https://monlamtech.github.io/Monlam-App-Store/>.

Ground Truth rather than recognising them with an × (i.e. as unknown), therefore requiring manual correction. The most effective strategy for improving model performance on such symbols is to focus on the training data, making sure it captures as wide a range of examples as possible, helping the model to learn diverse visual patterns and reduce misrecognition.

Symbol (MSS)	Symbol (Unicode)	Meaning
		symbols with no definite meaning, usually found at the beginning of a new section
	§	<i>gnyan</i>
		<i>mkhar</i>

Fig. 7 — Examples of symbols found in manuscripts from Leu 2011-30, alongside their transcription convention and meaning.

We also use ‘textual tags’ in Transkribus to mark up decorative features (see Fig. 8), alongside unclear text, additions, deletions, and text written in other scripts e.g. Zhangzhung Marchung (*zhang zhung smar chung*). These tags provide structured metadata that can be exported for analysis and aid in the systematic documentation and examination of scribal practices and codicological features. Tags can be enhanced with attributes, additional metadata that give information about the content of the tag, for example a description of a decorative feature, its colour, and so forth. Tags also help flag data that should be excluded from model training. In our case, we omit lines that contain ‘unclear’ tags used to mark up cases where text cannot be read due to damage,

poor image quality etc. Filtering out what our transcribers could not confidently transcribe reduces potential errors being introduced during training.

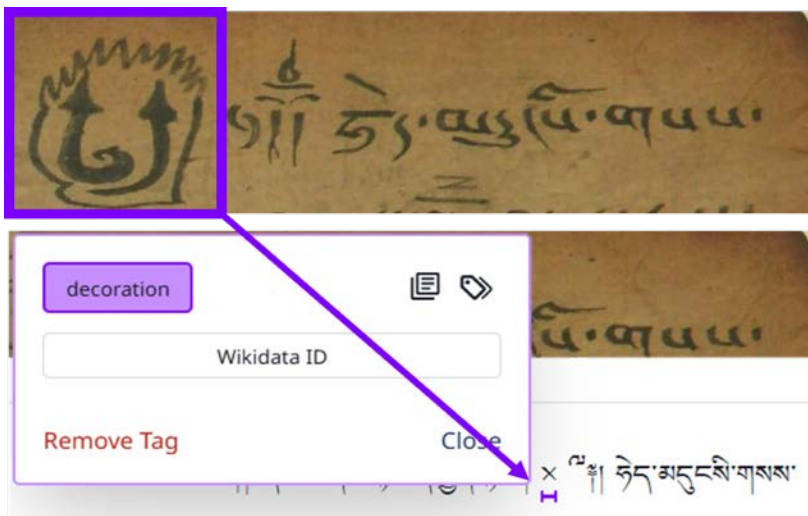


Fig. 8 — An example of a decorative feature tagged in Transkribus. Each tag is represented with a different colour. Folio taken from Leu 2011-30, vol 22, p.34.

Taken together, the issues outlined above show that our corpus poses challenges on multiple levels: its scale, the mixture of existing and newly collected materials, and the considerable palaeographical and codicological variation found within it. While we were able to anticipate and address some aspects of this complexity at the outset by developing a collaborative workflow and preparing training materials and manuals, an ongoing challenge has been the unpredictability inherent in working with a large and mostly unexplored corpus. Many challenges only presented themselves once the work was underway, requiring iterative adaptations to our approach, workflows and documentation. In the next section, we look at how our Layout Analysis and HTR models performed under these conditions.

3. Results and In-depth Error Analysis

The Layout Analysis and HTR models discussed in Section 2.1 above are trained to yield the best possible results when it comes to processing new materials. Because automatically transcribing difficult manuscripts is a challenging task (even for humans!), the results these models produce are unlikely to be 100% accurate. While small text corpora can accommodate minor errors because they are more amenable

to manual checking and correction before publication, this poses greater challenges for large electronic corpora. For some HTR projects, a lower baseline accuracy can be mitigated with LLM-based post-processing to normalise transcriptions and/or filter out noise. However, this approach is not viable for our corpus as it risks compromising diplomatic accuracy, which is crucial for in-depth philological and linguistic research. Instead, we focused on maintaining rigorous baseline standards and applying only targeted, rule-based corrections during post-processing. Consequently, we took the decision to exclude approximately 20,000 pages of transcriptions that were only double-checked, as they fell short of our Ground Truth requirements.

Whenever new models are trained, it is therefore important to be able to estimate how accurately they will recognise the text regions and lines and transcribe new images. In order to predict this accuracy, when training a model, a subset of images are kept 'hidden' (a so-called left-out test set)²⁵ so that once training is completed, the model's results can be compared with the Ground Truth and the differences can be calculated. The resulting evaluation metric for Layout Analysis models trained in Transkribus is called Loss on Validation Set with lower values indicating better performance.²⁶ We trained a custom Layout Analysis model, first on 7,212 pages, yielding a Loss of 4.8%. We then added new batches of images which had a rather different layout (i.e. both smaller and bigger pages, cropped in different ways, with added text in the margins and with more or fewer lines on one page). These new batches were more difficult to process, but the final custom Layout Analysis model ("Tibetan Manuscript 3", Transkribus model ID: 229409) trained on 15,006 images (130,869 lines) still yielded a Loss of 5.43%. This is higher than the previous model, but because of the wider variety of input, this last model is better at generalising, which means it is likely to yield better results for new, more varied data, which is what the majority of the PaganTibet corpus consists of.

For HTR models, the usual evaluation metric is called the Character Error Rate (CER), which gives an indication of how many characters are incorrectly transcribed per page; again, the lower the CER the better. CER can also be used to calculate differences between individual annotators who input or correct transcriptions. To test our general workflow, we measured this so-called 'Inter-Annotator Agreement' (IAA) on a subset of our data transcribed and corrected by the TNI monks, who were responsible for the first two rounds of correction (i.e.

²⁵ Note that the Transkribus platform refers to this as the 'validation' set, but it is crucially still left out of the training set so that it can be used for evaluation, error analysis and calculating Character Error Rates (CERs).

²⁶ According to Transkribus, a loss below 10% generally indicates reliable baseline detection. The same rule generally applies to CER, used to evaluate HTR models.

before the expert internal team members conducted a third and possibly fourth round of correction). The mean CER between annotators was 3.3% (median 2.0%), with 74.6% of lines showing a CER $\leq$ 5% - 30.3% achieving a CER of 0% - indicating high consistency overall in transcription practice. A small number of lines (5.4%), however, had a CER $\geq$ 10%, which prompted a more in-depth analysis of the disagreement between annotators. Common inconsistencies included whitespace handling (~60% of all disagreements) and character stacking issues; these also impacted the CER of our HTR models, addressed in more detail below.²⁷ In this section we discuss the results of the HTR models we trained and conduct an in-depth error analysis to measure the automatically calculated metrics, but also to see how well the models we trained perform on new parts of the corpus that exhibit new text topics and forms of umé script and how they compare to other existing models for transcribing Tibetan umé.

3.1 Automatic Evaluation Metrics for HTR Models

When comparing the accuracies of models, it is important to understand how much data they were given to train on, i.e. how big is the Ground Truth? This can be measured in pages or words, which are often used for HTR models developed for European languages that often have clearer word boundaries and images that show only one page. When working with Tibetan collections, however, due to the pothi format, one image often contains two or even three pages and there are no white spaces in the script to indicate word boundaries. For this and other reasons, calculating the number of lines a model was trained on can be a good alternative. In this section we will therefore report the number of lines the models were trained on, as well as an approximation of the number of pages.

We systematically compared the results of the seven different umé models we trained to check our progress at every iteration.²⁸ We then compared this to the other models that are publicly available for Tibetan umé HTR at the time of writing: two developed within the ERC-funded TibSchol project in Vienna, which are available on Transkribus, and another two developed by the BDRC, available through their

²⁷ See Griffiths & Meelen (forthcoming) for a more detailed discussion of the IAA analysis.

²⁸ The seven HTR models were trained between July 2024 and June 2026. These are publicly available on Transkribus: PaganTibet Ume 1 (ID: 131553), PaganTibet Ume 2 (193821), PaganTibet Ume 3 (328353), PaganTibet Ume 4 (428273), and PaganTibet Ume 5 (443525), PaganTibet Ume 6 (511497), and PaganTibet Ume 7 (579189).

custom-made app. Both the BDRC and TibSchol produced script-specific models for both Drutsa (*'bru tsha*) and Betsug (*dpe tshugs*) varieties. Since the texts in the PaganTibet corpus contain more Drutsa(-like) umé, we compared these to get the best possible results. Table 1 shows an overview of the Character Error Rates (CERs) reported for each of the PaganTibet (PT) HTR models for umé and the specific Drutsa umé models from the BDRC and TibSchol, alongside the number of lines and estimated number of pages they were trained on.

	CER	Training lines	Training pages (est.)
BDRC Drutsa	3.7 ²⁹	24,000 ³⁰	2,400
TibSchol Drutsa	1.4	3,803	422
PT-Ume1	6.99	2,338	400
PT-Ume2	5.56	3,952	900
PT-Ume3	6.26	5,064	1,200
PT-Ume4	4.57	8,630	1,800
PT-Ume5	4.87	10,238	2,100
PT-Ume6	4.54	12,687	2,500
PT-Ume7	3.71	16,361	3,200

Table 1 — Character Error Rates, number of lines and (estimated) pages to train the models.

Both the BDRC and TibSchol models have lower CERs than any of the PaganTibet models. Superficially, for the BDRC model this is entirely expected since it is based on a significantly larger set of training data

²⁹ This is the CER reported for the Drutsa model used for the BDRC custom OCR app (Elie Roux, p.c.) in October 2025. The app provides an offline, user-friendly interface and is publicly available (<https://github.com/buda-base/tibetan-ocr-app>). This specific BDRC Drutsa training data is also publicly available online on HuggingFace <https://huggingface.co/datasets/openpecha/OCR-Drutsa>. The BDRC are working on updating their models.

³⁰ The reported lines are actually >32,000, but only a subset of those were used for training. In addition, this number includes both lines and 'line fragments', i.e. part of lines, which is why the estimated number of pages for the BDRC model is about 10% of this number. Other models have line fragments too, but only for title pages (Elie Roux, p.c.).

(24,000 lines or 2,400 pages outnumber the majority of the other models). Based on CERs, the TibSchol Drutsa model clearly outperforms all others with a CER of 1.4%, based on just 3,803 lines (422 pages). In size, it is most comparable to the PT-Ume2 model when looking at the lines (which is the most accurate measure for comparing size), but the PT-Ume2’s CER is much higher at 5.56%. Apart from the size of the training set, it is also important to pay attention to the form and content. Both the BDRC and TibSchol models produce transcriptions in Wylie, while the PaganTibet models produce Tibetan Unicode. Due to differences in the display protocols of Tibetan Unicode stacks for different web browsers, apps and operating systems (Windows vs Mac) when it comes to non-standard stacks, special symbols, double vowels, etc. the error rates can end up being higher as it is not always easy to spot errors during data preparation or when manually correcting training data. Fig. 9 shows a range of examples of abbreviated forms with non-standard Tibetan Unicode stacks and how they are displayed on different operating systems. Double vowels in particular are almost impossible to see on Macs, making detailed error correction of diplomatic transcriptions that contain a high number of non-standard forms more challenging.

	Windows	Mac
		
		
		
		

Fig. 9 — Examples of how Windows and Mac operating systems differ when it comes to displaying Tibetan Unicode stacks that are non-standard (e.g. double vowels).

However, because of the collaborative nature of the PaganTibet workflow, using Wylie was not an option and all data used for training in-

put and model output was only done in Tibetan Unicode. Some resulting stacking issues are shown in Fig. 10. A large amount of the training data for the BDRC model was also originally in Tibetan Unicode, but then converted into Wylie after which CERs were calculated. Although good converters from Tibetan Unicode stacks to Wylie and vice versa are available (both online and built into the BDRC OCR app) issues can still arise.

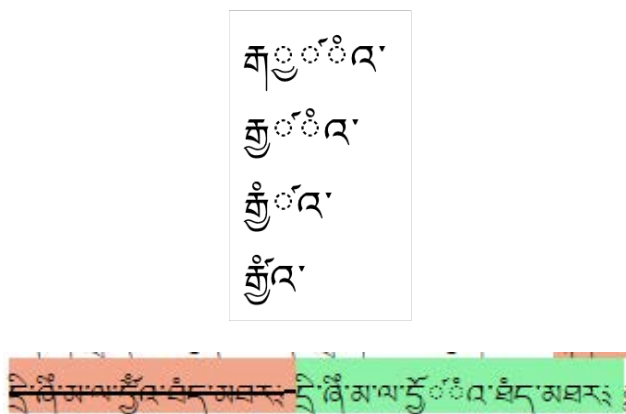


Fig. 10 — An example of stacking differences and how this can cause errors in HTR training and evaluation. Here, the model correctly reads the manuscript (red) but renders it differently from the Ground Truth (green).

Another reason why the Wylie-trained models achieve lower CERs is because of the more limited range of output options available. Fig. 11 shows the range of characters that are possible for the TibSchol Drutsa model compared to those for the most advanced PaganTibet model, PT-Ume5. With over twice as many options to choose from, it is not surprising that the TibSchol model's CER is much lower as the chance for the model to go wrong is much reduced. Looking at the detailed symbol lists in the training data also helps to identify errors and inconsistencies in the training data. In the TibSchol model, for example, three different types of apostrophes were used to transcribe the *achung* (འ), while in the PT-Ume5 model, both the capital (X), lower case (x) as well as the Tibetan Unicode × (U+0F1D) were used to transcribe blurry/illegible or otherwise unknown syllables. In subsequent training rounds these issues were corrected in the Ground Truth to improve the CERs and, potentially, the overall results when transcribing new data.

' * + . 0 1 2 3 4 5 6 7 9 : <ctc> <space> @ D I M
N O S T ^ a b c d e g h i j k l m n o p r s t
u w x y z | ß ´ ¨ ¯ ¸ º »

[illegible]

A final important reason why the CERs are generally slightly higher for the PaganTibet HTR models is related to the content of the source materials. Both the BDRC and TibSchol models are trained on Buddhist texts that are generally written in a rather standardised form of Tibetan, unlike the PaganTibet corpus (see Berounský & Ramble 2026). There are some abbreviations in these Buddhist texts too, but they are few in number and occur frequently in exactly the same form. This means that it is easier for the model to recognise them and transcribe them correctly. In addition, most HTR models are trained using Recurrent Neural Networks (RNNs) or similar architectures, which means they can take the preceding and following context of predicted characters into account as well. As a consequence, corpora with coherent topics and generally more-standardised language like those in the Buddhist training data of the BDRC and TibSchol models are easier to transcribe (both for humans and for HTR models) because words can be predicted from the context too, even if the glyphs in the manuscripts are unclear, blurry or abbreviated in any way. The unpredictable content, alongside the non-standard forms of the Tibetan language and scripts as described in Section 2.3 above, explain why higher CERs can be expected for the PaganTibet corpus since the CER calculation is done on more challenging test data.

Overall, measuring CERs can help to evaluate the accuracy of HTR models and they can serve as useful comparanda when developing new models, especially when the new data is expected to be similar in

form and content to the data the models were trained on. The accuracy of models and especially their capacity to generalise beyond the specific data they were trained on, can only truly be compared when evaluations are done on the same test sets. This should be borne in mind when developing general 'benchmarks' that are subsequently used for testing and comparing models. If these benchmarks contain mainly one type (genre, script style, image quality, etc.) of data, have different transcription conventions (e.g. for expanding abbreviations or not) or still contain errors, then models trained on a different type of data are likely to yield lower CERs. But these quantitative comparisons based on CERs do not always guarantee better results for more specific corpora in need of transcription or in general, however, which is why qualitative evaluations and in-depth error analyses of the output are important too.³¹ To illustrate this, in the next section we therefore present the results of the same models on some of the more challenging texts in the PaganTibet corpus we found so far.

3.2 Results and In-depth Error Analysis

A well-known issue for any type of computational model is 'overfitting'. When a model is overfitting, it generally means that the accuracy rates (CERs in the case of HTR modelling) are good, but that the model is 'fitted' too much on the data it trained on. As a result, it is unable to generalise and yield good results on new data that differs from the input. It is therefore useful to test any model's ability to generalise and go beyond the type of images found in the training data. Furthermore, since our main challenge is to create searchable eTexts for the PaganTibet collections and these collections are too big to transcribe or even correct manually, it is of the utmost importance to choose the right HTR model to transcribe the hundreds of thousands of pages in the best possible way.

In addition to the general quantitative CER analysis discussed in the previous section, we therefore also conducted several rounds of qualitative analysis - one of which we present here. To make this optimally useful, we deliberately chose test images from two separate Leu collections that differ significantly from the existing training data. The first set of five images (shown in Fig. 12) forms the start of a *bya rdang* ritual text found in the Leu 2016-10 collection. These images are challenging because they are not consistently cropped (the title page in particular is in a completely different format) and some pictures are out of focus and/or partially damaged.

³¹ See Meelen & Griffiths (2025b:10-12) for a case study on texts on Vampire subjugation (*sri gnon*) rituals.

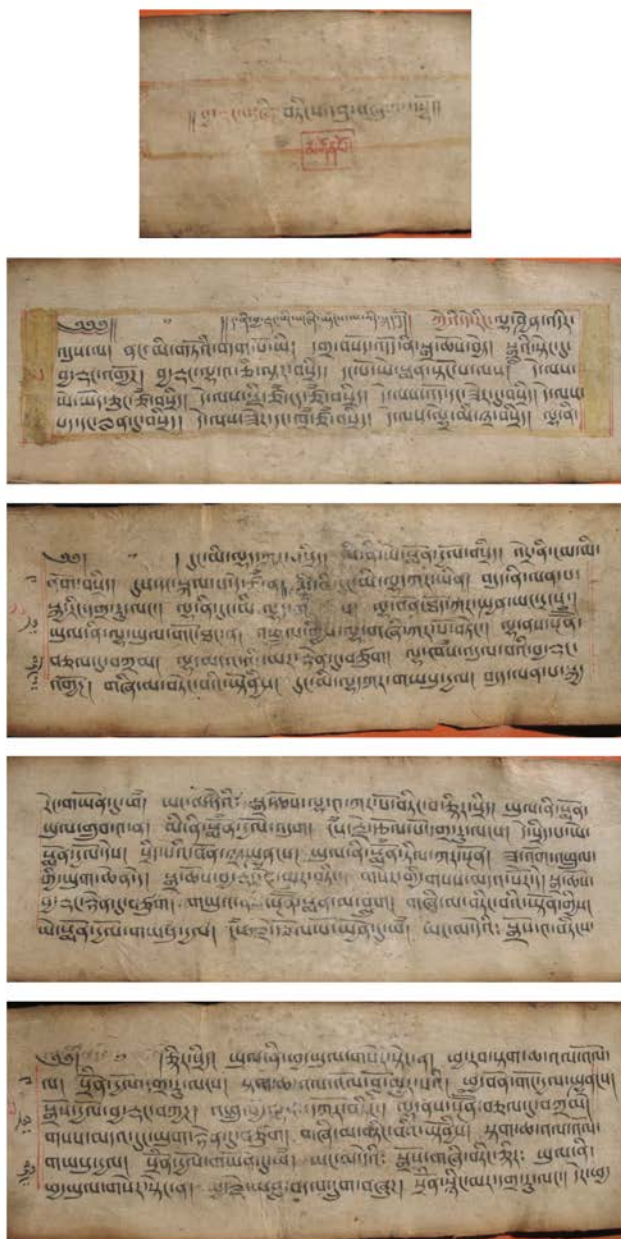


Fig. 12 — First five pages of the *Leu bya rdang* text in the Kalpa Bön collection (referred to as *Leu-Kalpa1-5* for the in-depth error analysis), *Leu 2016-10*, vol. 2, pp. 27-31.

The second set of five images (shown in Fig. 13) forms the start of a fox

fumigation (*wa bsang*) ritual described in Berounský (2019). Just like other images in the Leu 2011-30 collection, the content, script, non-standard abbreviations and symbols pose significant challenges.

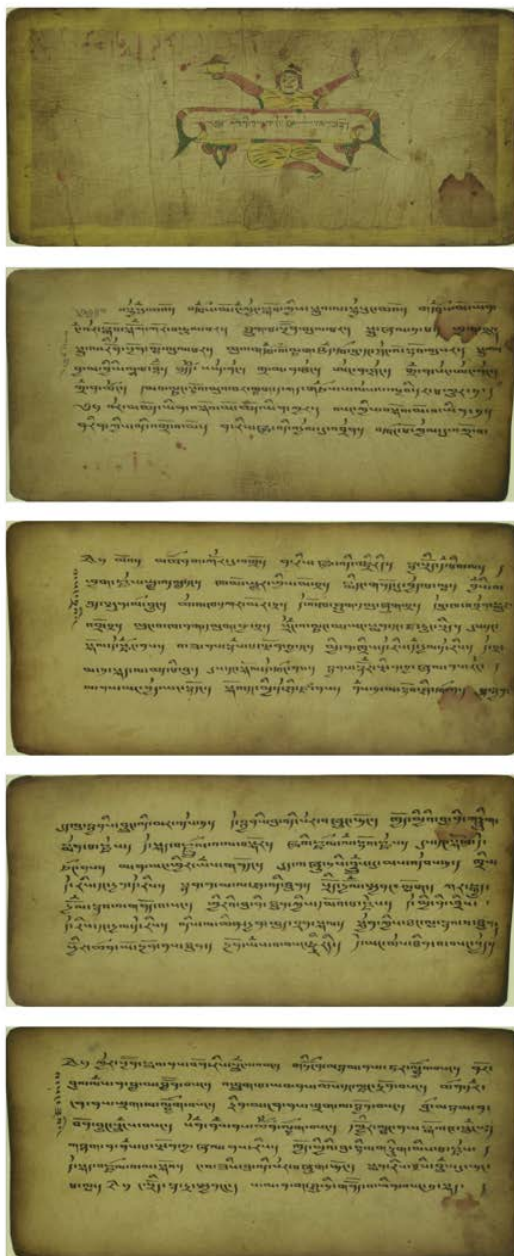


Fig. 13 — First five pages of the “Long-nosed Fox” in the Leu collection (referred to as Leu-vol26_1-5 in the error analysis), Leu 2011-30, vol. 26, pp. 1-5. © Gansu wen hua chu ban she.

The results of detailed error analyses for these ten test pages are shown in Fig. 14 and exact percentages of differences from manually triple-checked, Gold Standard transcriptions, counting all characters (including punctuation and spaces) that have been erroneously inserted, substituted or deleted are given in Table 2.

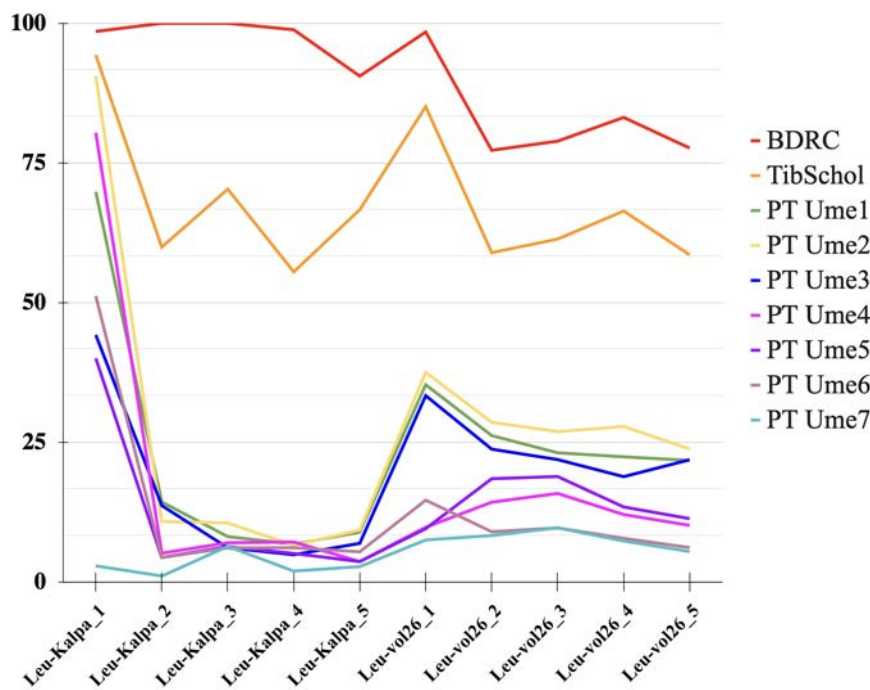


Fig. 14 — Percentage differences (errors) for each of the ten test pages comparing the BDRC, TibSchol and PaganTibet (PT) models.

Test Pages	BDRC	Tib Schol	PT Ume1	PT Ume2	PT Ume3	PT Ume4	PT Ume5	PT Ume6	PT Ume7
Leu-Kalpa_1	98.53	94.29	69.81	90.57	44.19	80.39	40	51.16	2.86
Leu-Kalpa_2	100	59.92	14.29	10.8	13.65	5.13	4.36	4.36	1.04
Leu-Kalpa_3	100	70.3	8.1	10.53	6.09	6.99	6.29	6.09	6.34

<i>Leu-Kalpa_4</i>	98.84	55.5	6.74	6.51	4.85	7.13	5.08	6.12	1.93
<i>Leu-Kalpa_5</i>	90.54	66.62	8.89	9.29	6.89	3.62	3.62	5.38	2.72
<i>Leu-vol26_1</i>	98.44	85.07	35.29	37.5	33.33	9.76	9.52	14.63	7.5
<i>Leu-vol26_2</i>	77.26	58.93	26.18	28.55	23.75	14.26	18.46	8.98	8.29
<i>Leu-vol26_3</i>	78.9	61.37	23.1	26.88	21.88	15.82	18.85	9.66	9.67
<i>Leu-vol26_4</i>	83.14	66.38	22.38	27.82	18.82	12.05	13.4	7.8	7.3
<i>Leu-vol26_5</i>	77.66	58.54	21.74	23.78	21.86	10.11	11.31	6.13	5.42
Average % errors	90.33	67.69	23.65	27.22	19.53	16.53	13.09	12.03	5.31
<i>Model CER</i>	3.7	1.4	6.99	5.56	6.26	4.57	4.87	4.54	3.71
<i># of lines in training data</i>	24,000	3,803	2,338	3,952	5,064	8,630	10,238	12,687	16,361

Table 2 — Percentage differences (errors) for each of the ten test pages comparing the BDRC, TibSchol and PaganTibet (PT) models with CER and number of lines in training data for reference.

One immediate observation we can make is that lower CERs do not correspond to better results on these challenging test pages from the PaganTibet corpus. In fact, we observe almost the opposite pattern with the BDRC and TibSchol models producing over 90% and 67% errors on average respectively, while all PaganTibet models (even the earlier ones with higher CERs trained on fewer lines of data) range between 5-28% errors on average. The deliberate choice of varied input allows us to analyse results in more detail: blurry vs non-blurry, different text colours, title vs full page, etc. as shown in the title pages of both test sets. The TibSchol model clearly struggles with the blurry title page of the Kalpa Leu text (>94% errors), where some PaganTibet models perform much better (PT Ume7 only has 2.86%). Other PaganTibet models also struggle with the blurry title page, which is no doubt due to the fact that there were no or very few blurry pages in

the training data. Once a model trains on blurry or damaged images, it ‘learns’ how to deal with these and performs better in future cases. Larger numbers of lines in the training data also help with this as blurry and damaged pages typically require better input from the context and context is easier learnt from larger sets of training data.

Another clear result is the effects of different hands, writing styles and/or umé varieties. All the pages from the second set (volume 26 of the 30-volume Leu collection) are written in a very different style from the umé found in the training data of most models. For both the BDRC and TibSchol models, where error rates are over 60% across the test set, this makes little difference, but the first three (much smaller) PaganTibet umé models clearly struggle more with Leu-vol26. PT Ume4 and Ume5 are significantly better at transcribing more varied and more challenging input, while PT Ume1, 2 and 3 did not have as much challenging data in the training set, which is why error rates for these are higher. PT Ume4 and Ume5 are also trained on more data and both are better at generalising, i.e. they still get good results (lower error rates) on pages that are completely unfamiliar in terms of script styles/handwriting etc. It is not until we start adding a small amount of training lines similar to the Leu30 collection for PT Ume6 and Ume7 that error rates for those test pages go down. Finally, when comparing individual pages, we see that some PT models appear to perform worse than earlier versions on the same page, such as images 3 and 4 in the Kalpa Leu *bya rdang* text and image 1 from Leu-26. This is likely due to a combination of page-specific challenges - particularly damaged sections where performance is more dependent on chance - and the natural variance introduced by training dynamics, which means successive iterations will not uniformly improve on every test case even when overall CER improves.

When it comes to comparing three different models (TibSchol, BDRC and PT) directly we notice that although the CER of the PT Ume5 model is slightly higher than that of PT Ume4, it still outperforms Ume4 on average, suggesting that it is better at dealing with different and more challenging new data. Fig. 15 shows an example of a direct comparison for one of the Leu vol-26 pages with the specific errors in the output highlighted in red (deletions) and green (additions). This allows us to see what specific errors each model makes. For example, *yig mgo* and other punctuation and ornamental features were clearly not part of the TibSchol or BDRC training data as both models produce various errors with these while the PT Ume7 model transcribes these correctly most of the time. Something all models struggle with are symbols such as those described in Section 2.3 above, here transcribed as “X” in line 4. The reasons for TibSchol's worse performance on this page (and the rest of this test set) are mainly related to

the fact that it lacks varied input as it was trained on much more coherent Buddhist texts that do not exhibit the orthographic peculiarities we find in the Leu collections. This more homogeneous input makes it harder for the model to generalise to other hands, genres, linguistic and orthographical features and different types of umé. The BDRC model faces similar issues as it was partly trained on the same input, but despite the much larger number of lines in the training data, the model performs much worse on the PaganTibet test pages. One important reason for this lies in the fact that the input was of a mixed nature (both Wylie and Tibetan Unicode stacks), which was automatically converted to create a model that outputs Wylie, but is then automatically converted back to Tibetan Unicode. As is clear from Fig. 15 below, this conversion does not always go right, with fragments of Wylie left in between the Tibetan unicode stacks.

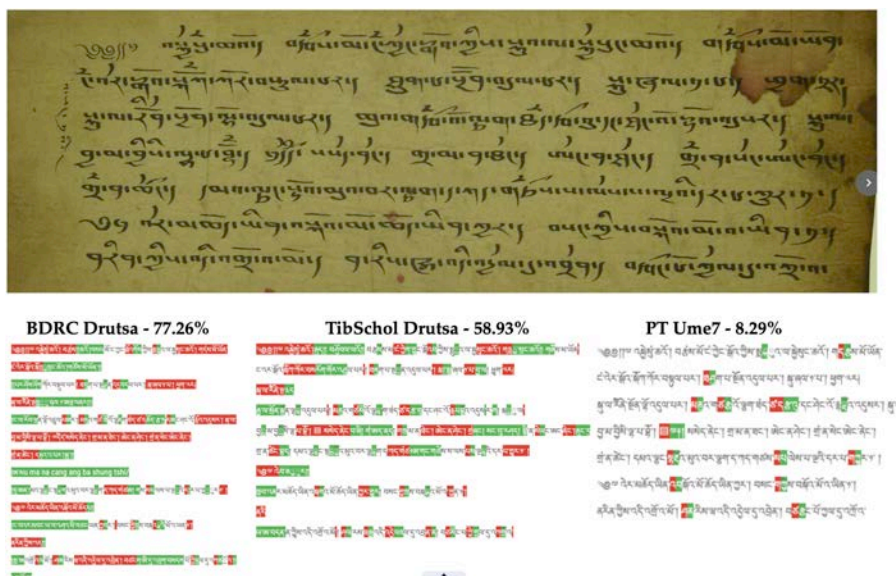


Fig. 15 — Visualisation of differences (deletions in red and additions in green) between the HTR models' output and the Gold Standard to give an impression of overall errors for p. 2 of the Leu-vol26 where TibSchol had 58.93%, BDRC 77.26% and Ume7 8.29% errors.

Another major difference between the BDRC training data (comprising 24,000 lines) and that of PaganTibet (comprising at most 16,361 lines for PT Ume7), is the rounds of manual correction of the original training data. The PaganTibet workflow described in Section 2.1 above consisted of at least three rounds of manual correction (with a fourth and final check for a subset of images), while the BDRC likely utilised fewer rounds of checks. Although this tripled the time spent on the

preparation of Ground Truth, it was deemed necessary due to the difficult nature of the material and even after three rounds of manual correction, more errors were found and corrected in the validation set (a small amount of ‘unseen’ data given to the model during training to compare the model’s predictions alongside the Ground Truth) of PT Ume4, Ume5 and Ume 6 before the next models were trained. These multiple checks by experts are extremely time-consuming,³² but the results show that with challenging, non-standard collections, multiple checks are the only way forward to gain reasonable results. Even if internally calculated CERs are low and therefore look promising, HTR models that yield far over 50% error rates on test data that is representative of the rest of the corpus are not useful since they would require an enormous amount of time to do further rounds of correction after the automatic transcriptions are done.

There are some overall lessons learnt from both the quantitative and the in-depth qualitative analyses presented here. First, CERs can be useful benchmarks when training new models for datasets that are homogeneous and coherent in both form and content: the TibSchol model with the lowest CER of 1.4% is likely to perform well on other manuscript images of a similar nature, both in terms of script type as well as genre, especially when Wylie transcriptions are required. Furthermore, performance generally improves when more training data is added, as shown by the incrementally lower averages of errors on the challenging test set cases for the various PT Ume models. But more data is not necessarily better for either CERs or for the challenging test cases as the largest HTR model from the BDRC shows both a higher CER than TibSchol and a much higher error rate on the PaganTibet test set. This confirms the results also found by Meelen & Griffiths (2025b) comparing yet a different type of umé pages as a test set. It shows that careful selection of ideally varied training data, combined with triple-checking (at least) of every line by expert annotators who are well-trained and follow a consistent set of protocols for diplomatic transcriptions³³ is of the utmost importance when dealing with large, varied corpora containing texts of various genres with non-standard linguistic and orthographical features. Finally, HTR postprocessing can help to filter out frequent errors to prepare corpora for subsequent stages of training and downstream NLP tasks, such as turning diplomatic transcriptions to normalised eTexts and adding (linguistic) annotation.

³² It took a team of ten annotators who worked part-time over a year to double-check 30,000 pages and another team of experts several months to triple-check just over 3,000 of those.

³³ Such as those outlined in the PaganTibet HTR manual (Meelen & Griffiths 2025a) and cheatsheet (Griffiths & Meelen 2025a).

4. Conclusion

The aim of this article was to illustrate how to get from manuscript to eText, using the PaganTibet corpus as a case study in large-scale text processing. Its sheer size, the heterogeneity of manuscripts, and the considerable variation in script, alongside other challenges such as abbreviations and symbols, call for a carefully structured, collaborative approach. As our workflow demonstrates, drawing on combined expertise can successfully navigate many of these challenges. Ultimately, our methods and models demonstrate that creating searchable eTexts for a large and largely unexplored corpus is feasible, but that carefully planning workflows and taking into account the potential challenges described in Section 2 is essential.

Our results show that while conventional evaluation metrics such as CERs are informative, they do not always reflect the practical performance of HTR models. In the case of the PaganTibet corpus, models trained specifically on diverse and representative input, even those with slightly higher CERs consistently outperformed models trained on larger but more homogeneous datasets when applied to novel and complex materials. Conducting in-depth qualitative analyses proves equally useful, as it flags sticking points such as the angle of the *'greng bu* in the Leu vol. 26 sample, which we then incorporated into future rounds of training. It also highlights Unicode symbols that have found their way into our Ground Truth, but that should not be there, e.g. different types of X symbols. With the latter, we can update our HTR postprocessing script to search and replace these symbols to further enhance existing transcriptions. These two steps enabled us to create a 7th model, which surpassed the earlier iterations and has become our primary solution for generating automatic diplomatic transcriptions of the remaining corpus. For further processing, in addition to diplomatic transcriptions that reflect all non-standard spellings and symbols, we will create normalised versions of the eTexts (see Griffiths & Meelen 2025b and Meelen & Griffiths 2026). In these normalised eTexts, abbreviations are expanded and the orthography is standardised to make the transcriptions easily searchable and ready for additional annotation (e.g. to show linguistic and content features).

Bibliography

*An.

- Leu 2011-30 = 'Brug thar, ed. 2011. *mDo smad mda' tshang yul gyi gna' dpe phyogs bsduṣ mthong ba don ldan*. 30 vols. Lanzhou: Gansu wenhua chubanshe 甘肃文化出版社 [Gansu Culture Publishing

- House]. BDRC MW2PD17384.
- Leu 2013-60 = 'Brug thar, ed. 2013. *mDo khams yul gyi bod yig gna' dpe phyogs bsdus mthong ba 'dzum bzhad*. 60 vols. Lanzhou: Gansu wenhua chubanshe 甘肃文化出版社 [Gansu Culture Publishing House]. BDRC MW2PD17385.
 - Leu 2016-10 = Thar, Tsering, and Ngawang Gyatso, eds. 2016. *gNa' rabs bon gyi dpe dkon bris ma*. 10 vols. Xining: mTsho sngon mi rigs dpe skrun khang.
 - Leu 2024-100 = Luosai 洛赛, Liang Jixiao 梁吉效, and Quan Yongkang 全永康. 2018, 2020, 2021, 2024. *'Brug chu'i dmangs khrod gna' rabs bon gyi dpe dkon bris ma phyogs bsgrigs*. 100 vols. Lanzhou: Kan su'u rig gnas dpe skrun khang.
 - *dGa' thang 'bum pa nas bon gyi gna' dpe* = Pa tshab Pa sangs dbang 'dus and Glang ru Nor bu tshe ring, eds. 2007. *Gtam shul dga' thang 'bum pa che nas gsar rnyed byung ba'i bon gyi gna' dpe bdams bsgrigs*. Lhasa: Bod ljongs dpe rnying dpe skrun khang.
- Berounský, Daniel**
 "Burning the incestuous fox. A Tibetan fumigation ritual (wa bsang)", *Études mongoles et sibériennes, centrasiatiques et tibétaines*, 50, 2019, pp. 1–32. doi:10.4000/emscat.3782.
- Berounský, Daniel & Charles Ramble**
 Leu Texts from Eastern Tibet: The Emergence of a Canon of Pagan Tibetan Literature? In *Canons, Kangyurs, and Collections – Multidisciplinary Approaches in the Study of Tibetan Canonical Literature*, edited by Markus Viehbeck, Bruno Lainé, Kurt Tropper, and Verena Widorn, Verlag der Österreichischen Akademie der Wissenschaften; Beiträge zur Kultur- und Geistesgeschichte Asiens. 2026.
- Chirkova, Katia, Rolando Coto-Solano, Rachael M. Griffiths & Marieke Meelen**
 "Comparing efficacy of IPA vs Pinyin romanisation transcriptions for complex tonal languages: A case study in Baima" in *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Honolulu, Hawai'i, USA, Association for Computational Linguistics. 2025. pp. 170-181.
<https://aclanthology.org/2025.computel-main.20/>
- Erhard, Franz Xaver**
 "Text and Layout Recognition for Tibetan Newspapers with Transkribus", *Revue d'Etudes Tibétaines*, no. 74, 2025, pp. 128-171.
- Griffiths, Rachael M.**
 "Handwritten text recognition (HTR) for Tibetan manuscripts in cursive script", *Revue d'Etudes Tibétaines*, no. 72, 2024, pp. 43–51. doi: 10.1553/TibSchol_ERC_HTR

Griffiths, Rachael M. & Marieke Meelen

HTR Input & Correction Cheat Sheet: 10 Basic Rules and Protocols for Diplomatic Transcription. Zenodo. 2025a. doi: 10.5281/zenodo.17251317.

Griffiths, Rachael M. & Marieke Meelen

How 'Pagan' Is My Text? Information Extraction from Untranscribed Data. In *Anthology of Computers and the Humanities*, 3. 2025b. pp.1262–1273. doi: 10.63744/ayiz0ulyis4f.

Griffiths, Rachael M. & Marieke Meelen

Annotation quality and error typology in low-resource Tibetan HTR (forthcoming).

Karmay, Samten G.

A general introduction to the history and doctrines of Bon. In *The Arrow and the Spindle*, Kathmandu: Mandala Book Point, 1998 [1975]. pp. 104–156.

Kværne, Per

The Bon Religion of Tibet. The Iconography of a Living Tradition. London: Serindia Publications. 1995.

La Xian 拉先

Baima Zangzu zongjiao wenhua yanjiu 白马藏族宗教文化研究 [Recherche sur la culture religieuse des Tibétains de Baima], Zhongguo Zangxue Chubanshe. 2022.

Meelen, Marieke & Rachael M. Griffiths

HTR Input and Correction Manual. Zenodo. 2025a. doi:10.5281/zenodo.17257009.

Meelen, Marieke & Rachael M. Griffiths

"Collaborative Workflows for Handwritten Text Recognition in Under-Resourced Manuscript Collections", *Journal of Open Humanities Data*, vol. 11(1-54). 2025b. doi: 10.5334/johd.388.

Meelen, Marieke & Rachael M. Griffiths

"Historical Tibetan Normalisation: rule-based vs neural & n-gram LM methods for extremely low-resource languages", *Proceedings of the AI4CHIEF conference* (2026), Springer.

Meelen, Marieke, Sebastian Nehrdich, & Kurt Keutzer

"Breakthroughs in Tibetan NLP and Digital Humanities", *Revue d'Etudes Tibétaines*, no. 72, 2024, pp. 5–25.

Ngawang Gyatso

mDo smad kan lho yul du bon bstan dar ba'i lo rgyus rags rim tsam brjod pa. In *Mdo smad zhib 'jug*, vol. 1, edited by Ting 'dzin bkra shis, Beijing: Mi rigs dpe skrun khang, Kan su'u bod kyi shes rig zhib 'jug khang. 2005. 174–81.

Ramble, Charles

"The Classification of Territorial Divinities in Pagan and Buddhist Rituals of South Mustang." In *Tibetan Mountain Deities, their Cults and Representations*, edited by A.-M. Blondeau, Vienna: Verlag der Österreichischen Akademie der Wissenschaften. 1998. pp. 123–43.

Sabbagh, Christina, Franz X. Erhard, Robert Barnett & Nathan Hill

Divergent Discourses Custom Image Preprocessing (Forked Binarisation). Zenodo. 2024. doi: 10.5281/zenodo.14523007.

Sabbagh, Christina

"Enhanced HTR Accuracy for Tibetan Historical Texts - Optimising Image Pre-processing for Improved Transcription Quality", *Revue d'Etudes Tibétaines*, no. 74, 2025, pp. 81–127.

Snellgrove, David

The Nine Ways of Bon, Oxford: Oxford University Press. 1967.

Werner, Eric and Markus Viehbeck

"Automated Character Recognition in Early Tibetan Canonical Literature", *Revue d'Etudes Tibétaines*, no. 79, 2025, pp. 55–65.

