Text and Layout Recognition for Tibetan Newspapers with Transkribus

Franz Xaver Erhard

(Leipzig University)

Digitising and, in particular, transcribing, i.e., performing automatic text recognition (ATR) on Tibetan language texts, despite the several efforts undertaken since the 1990s, remained unavailable for too long and large financial and human resources went into manually keying-in Tibetan texts. This technological lacuna hampered the development of Tibetan digital humanities, and the Tibetan studies community had to contend with the production of digital artefacts, usually PDFs of original texts, without being able to apply higher-level digital tools such as full-text search, not to speak of natural language processing (NLP). Recent developments in Artificial Intelligence made enormous advances in handwritten text recognition (HTR) possible, now allowing the training of HTR models for specific handwriting or print independently from the language.[1] The UK-German collaborative research project Divergent Discourses. Narrative Construction in Tibet, 1955–1962[2] takes advantage of these developments.

[1] The best-known approach to ATR is optical character recognition (OCR), which attempts to detect each individual character in a given text. Another recently more popular approach to ATR is handwritten text recognition (HTR), which looks at a whole line of text. HTR was developed specifically for writing with inconsistent font, as prototypically found in handwritten letters, ledgers and diaries. For a helpful comparison of the approaches, see Nockels *et al.* 2024.

[2] The project received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (https://gepris.dfg.de/gepris/projekt/508232945?

Erhard, Franz Xaver, "Text and Information Extraction for Modern (post-1950) Tibetan Newspapers and Print Publications with Transkribus", *Revue d'Etudes Tibétaines*, no. 74, February 2025, pp. 128–171.

Divergent Discourses aims to study the construction of narratives in Tibet in the mid-20th century, a crucial period of social and political change. To this end, the project studies narratives about historical events at the time of their origin and tracks their evolution over time. To do this, the project needs to build a digital corpus of Tibetan newspapers of that period as a basis for its analysis. As a first step, we have brought together available collections of Tibetan newspapers from seven different libraries across Europe, the US and India, thus building a corpus of 16 newspapers and almost 17,000 pages (see Erhard 2025 in this issue). The following steps in the workflow include digitising newspapers and applying custom-trained HTR models to extract information, including full e-texts, for further analysis.

1 *Digitisation and Text Recognition for Tibetan Newspapers*

Earlier research projects digitising Tibetan language newspapers generally focussed on the *Tibet Mirror* (*yul phyogs so so'i gsar 'gyur me long*) of which, over the past decades, more extensive collections came to light in various libraries in the United States, the United Kingdom, France, Germany, Austria, and India. Building on Paul Hacket's work, Columbia University pioneered[3] this effort and, in cooperation with the Beinecke Rare Books and Manuscript Library at Yale[4], the Musée Guimet, and the Collège de France, scanned their holdings of the *Tibet Mirror* and made them available freely between 2009 to 2013.[5] This

language=en), and from the Arts and Humanities Research Council (AHRC) under project reference AH/X001504/1 (https://gtr.ukri.org/projects?ref=AH%2FX001504%2F1). For more information on Divergent Discourses, see https://research.uni-leipzig.de/diverge/.

[3] For more information about the project led by Lauran Hartley from 2009 to 2013, see https://library.columbia.edu/libraries/eastasian/special_collections/tibetan-rare-books---special-collections/tharchin.html (accessed September 15, 2024).

[4] See https://beinecke.library.yale.edu/collections/highlights/tibet-mirror (accessed September 15, 2024).

[5] A table of contents and images are available https://openlibrary.org/works/OL17161360W/Yul_phyogs_so_so%CA%BEi_gsar_%CA%BEgyur_me_lon%CC%

project triggered substantial research on the *Tibet Mirror*; however, further digitising efforts to explore the newspaper's content in greater depth largely did not materialise.

Pavel Gorkhovskiy from Saint Petersburg State University (Moskaleva & Gorkhovskiy 2018), as well as a small team at the Collège de France, started to digitise the publication, including transcriptions, and make its content accessible (Wang-Toutain 2018). These efforts resulted in the newspaper's publication in the form of International Image Interoperability Framework (IIIF) manifestos[6] alongside some annotated manual transcriptions.[7] However, the relevant Digital Humanities tools, at least for the Tibetan language, are still lacking for a deeper exploration of the newspaper sources. Yet, effective tools for text recognition and information extraction are crucial for thoroughly exploring Tibetan newspaper corpora.

Text recognition alone is a significant step in the digitisation process. Still, more information would help create useful e-texts and e-corpora. For example, besides extracting the sentence “ཧྐ་ཁྲོན་ཀུན་རྗེས་ཡང་ཡུ་འདེབས་ཀྱིན་ཡོད་པ། །”, it would be helpful to know that the text string is a chapter heading. Being able to automatically identify structural elements such as headings, sub-headings, or page numbers subsequently allows for the automatic structuring of the e-text. Information extraction then goes beyond mere text extraction. It includes identifying and extracting structural information and, in later stages, entities and relations – although not relevant in the context of our project.

Recognising and identifying layout components, such as page numbers, columns, headings, captions, images, and illustrations, is

87_%28Tibet_Mirror%29?edition=key%3A/books/OL25732003M (accessed September 15, 2024) or https://archive.org/details/ldpd_6981643_000 (accessed September 15, 2024).

[6] The IIIF standard allows for easy and rich access, sharing, and embedding of images. For details, see the IIIF consortium’s webpage https://iiif.io/

[7] See https://salamandre.college-de-france.fr/archives-en-ligne/ead.html?id=FR075CDF_000IET002&c=FR075CDF_000IET002_de-310 (accessed September 15, 2024) F. Wang-Toutain is currently finalising her analysis of ornamental and illustrative parts of the *Tibet Mirror*.

crucial for text recognition and extracting information from historical newspaper sources. For example, to produce a readable e-text of a multi-column newspaper, the model needs to recognise the columns to maintain the correct reading order of the lines.[8] We will first focus on Tibetan text recognition before dealing with layout analysis and information extraction.

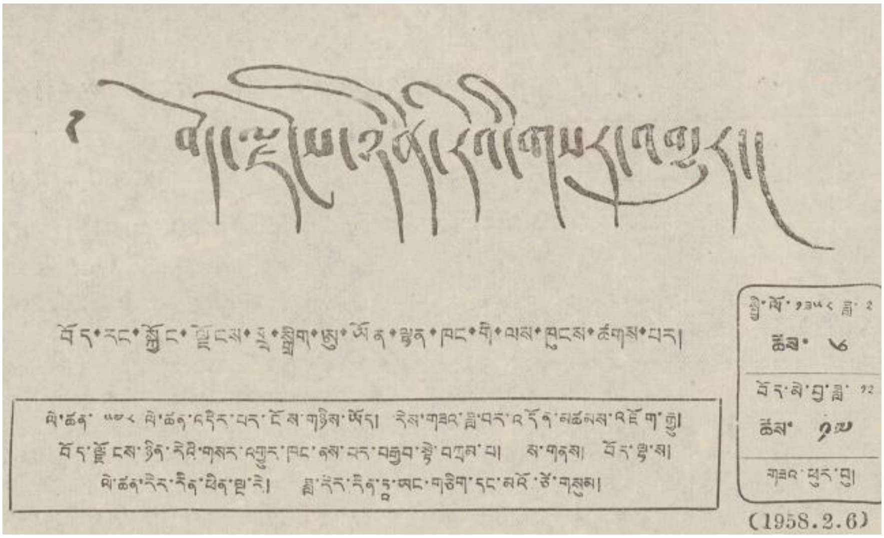
2 *Tibetan Automatic Text Recognition*

Tibet has a long literary history, with the earliest sources dating to the second half of the 7th century. The earliest Tibetan printed text produced from woodblocks that has survived is a 12th-century prayer book.[9] Tibetan literature has since developed into "one of the great literary traditions of Asia" (Cabezón & Jackson 1996: 11). It is written in Tibetan script in two general varieties: *uchen* (*dbu can*), literally meaning "headed letters" and *ume* (*dbu med*), literally "letters without a head" (Schubert 1950: 281). Printed text generally appears in the regular *uchen* typeface, while cursive *ume* scripts are used only in manuscripts. Still, in the 20th century, cursive scripts also appeared in print publications, particularly in early cyclostyle newspapers,[10] or later in the ornamental titles of many newspapers (Fig. 1), or as a decorative style for headlines and headings. However, the differences

[8] If the column layout is not recognised, most OCR or HTR models would read from left to right, jumping from one column to the next and, hence, producing nonsensical text. A solution to this problem is discussed in section 4 below.

[9] This early print produced on Tangut paper belongs to a collection from Khara Khoto and is preserved in the Institute of Oriental Manuscripts (IOM) in St Petersburg (Bradburne 1993: 278).

[10] The *Tibet Mirror* features a broad range of styles and scripts and occasionally whole pages are (hand-)written in cursive script and reproduced with the single drum RENO Steno duplicator in 1925 (Fader 2004: 258) or later from 1927 with a Double Crown lithographic press (Fader 2004: 334) and a newer demi-size litho hand press from 1934 (Fader 2009: 86). In the 1950s, the "News in Brief" (*Gsar 'gyur mdor bsdus*) published in Lhasa from 1953 to 1956 was produced in cursive U-me until April 1955 (Schubert 1958: 6).



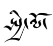
བོད་རང་སྐྱོང་ལྗོངས་གྲ་སྒྲིག་ཨུ་ཡོན་ལྷན་ཁང་གི་ལས་ཁུངས་ཚགས་པར།

ལེ་ཚན་ ༥༥༨ ལེ་ཚན་འདིར་པར་ངོས་གཉིས་ཡོད། རེས་གཟའ་ཟླ་བར་འདོན་མཚམས་འཛིན་ཀྱི།
བོད་ལྗོངས་ཉིན་རེའི་གསར་འགྱུར་ཁང་ནས་པར་བརྒྱབ་སྤེ་བཀྲམ་པ། ས་གནས། བོད་ལྷ་ས།
ལེ་ཚན་རེར་རིན་ཞིན་ཕྱེ་རེ། ཟླ་རེར་རིན་ཏ་ཡང་གཅིག་དང་མའོ་ཙེ་གསུམ།

སྤྱི་ལོ་ ༡༩༥༨ ཟླ་ ༢
ཚེས་ ༦
བོད་མེ་བྱ་ཟླ་ ༡༢
ཚེས་ ༡༨
གཟའ་ཕུར་བུ།

(1958.2.6)

Figure 1 Masthead of the Tibet Daily *(TID) February 6, 1958 with handwritten title by the 14th Dalai Lama (Hartley 2003: 87). (Oriental Institute, CAS, Prague XIV91-1959)*

between *uchen* and *ume* can go beyond mere style and outward appearance and affect the graphical representation of letters. For example, the term *spyi lo* (roughly "western year") is written སྤྱི་ལོ་ in standard *uchen* and  in the *drutsa* (*'bru tsha*) variation of *ume* as used in Figure 2.

2.1 Tibetan ATR: State of the Art and Challenges

Together with Tibetan's clustered orthography, where letters change their appearance when joined in ligatures, the writing conventions, despite the wealth of Tibetan literature, have made the development of Tibetan Optical Character Recognition (OCR) challenging.[11]

[11] Rowinski and Keutzer (2016) describe past research into Tibetan OCR, including their Namsel system. More recently, Google has made significant progress in the development of Tibetan OCR; see https://digital Tibetan.github.io/DigitalTibetan/docs/tibetan_ocr.html?highlight=ocr for an overview. Recently, the Norbu Ketaka project used Google's Tibetan OCR and enhanced it with further post-processing

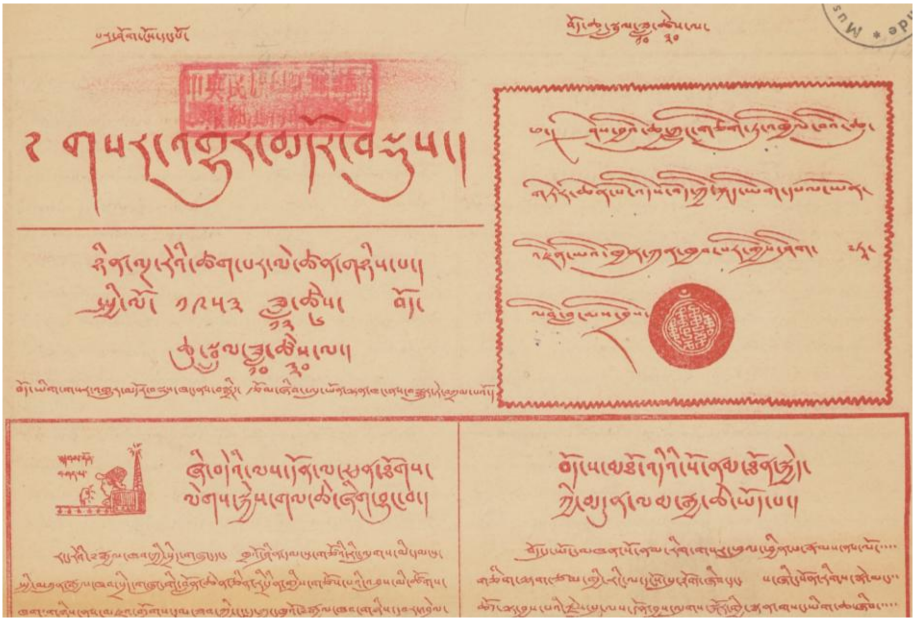


Figure 2 Example of different scripts in the News in Brief (NIB) from December 6, 1953 (Grassi Museum für Völkerkunde ASZAG9)

A significant challenge to OCR arises from the corpus's multilingual nature, which contains a significant share of English passages and horizontal and vertical Chinese. But more importantly, the various printing technologies used to produce Tibetan newspapers are usually limited to a specific script style or set of styles, including a variety of cursive scripts and a particular font.

While traditional print techniques prevailed for producing *pecha* (*dpe cha*), i.e., traditional woodblock prints, lithographic printing presses allowed cursive script, usually reserved for manuscripts, in printed mass media. The earliest Tibetan language newspaper, the *La dvags kyi ag bar* (*Ladakh Akhbar*), published by Moravian missionaries in the Western Himalayas, was produced using a simple duplicator or lithographic press. It sparked the evolution of other forms of publishing in Tibetan, in particular Tibetan publications on

steps, see Luo and van der Kuijp 2024. For the needs of the Divergent Discourses project, these approaches did not yield sufficiently accurate results.

lithographic print (*rdo par*), such as that used in the first newspaper published by the last Qing Amban in Lhasa in 1907, the *Tibetan Vernacular News* (*Bod kyi phal skad kyi gsar 'gyur*).

Tibetan movable type, interestingly, has been available in Western publications since the 19th century; however, probably due to their high cost, Tibetan types were not readily available in India. The Moravians, for example, published less prestigious materials with their local cyclostyle. Still, valued publications such as the Tibetan translation of the gospel of Matthew were printed at Unger Bros. (Th. Grimm) in Berlin in the early 20th century. Of high aesthetic quality, this so-called Jäschke type, although widely used in publications outside Tibet (Schubert 1950: 291–295), did not leave a great impression on the history of Tibetan typefaces.

In the newly founded Republic of China, the Mongolian and Tibetan Affairs Bureau (MTAB) lithographically published the bilingual *Bod yig kyi phal skad kyi gsar 'gyur* (*Zangwen baihuabao,* "Tibetan Language Vernacular News") in Peking. In 1915, the MTBA hoped to transfer the monthly newspaper to production in movable type (*lcags par, lcags 'bru*). However, MTBA was, for unknown reasons, unable to relaunch the Tibetan language edition, and the newspaper was discontinued.[12]

The best-known Tibetan newspaper published in India, the *Tibet Mirror*, used lithographic printing technology from its first edition in 1925 until publication ceased in 1963.[13] The first newspaper published by the People's Liberation Army (PLA) in Lhasa after the annexation of Tibet, the *Gsar 'gyur mdor bsdus* ("News in Brief"), published in

[12] See Erhard & Hou 2018: 10. See also the discussion in Pistorius 2019: 11–12.

[13] Tibetan language publications in India before the arrival of the 14th Dalai Lama in exile in 1959 were primarily dominated by Christian missionaries. Tharchin's Tibet Mirror Press was, for longer periods, funded by the Scottish Catholic Mission, and the remainder of Tibetan language publications were published by Moravians. They worked intensively on the Tibetan translation of the Bible, language primers and other publications that aided them in their proselytising activities. Interestingly, some of these early 20th-century publications were printed outside Tibet by, e.g. Gebr. Unger in Berlin, Germany. On the history of Tibetan moveable type, see Schubert 1950, in particular, pp. 292–295.

Lhasa from 1953 to 1956,[14] was initially lithographically produced in Tibetan handwritten cursive *drutsa* with an irregular single- or two-column layout. Until lead typesetting with movable types and offset printing were introduced in newspaper production in Lhasa in May 1955, newspapers featured various irregularities in manuscripts, such as abbreviations or variations of the scribal hand.

Mass-produced media had been established in other Tibetan areas of the young People's Republic of China (PRC) already a few years earlier, with the *Mtsho sngon bod yig gsar 'gyur* ("Qinghai Tibetan Language News") starting in 1951 [15] being the earliest Tibetan language newspaper in Communist China. The paper was printed in movable type from its inception. In the first half of the 1950s, thus, a movable type for Tibetan emerged that was widely used in newspapers but also in the now evermore frequent books published by the recently founded, state-run Minorities or Nationalities Publishing Houses (*mi rigs/mi dmangs dpe skrun khang*).[16] The next fundamental change in printing took place only with the introduction of computers, probably in the 1990s. Until then, the appearance of most print publications remained largely the same (Erhard 2018: 117–118).

2.2 *Previous research, approaches and limitations*

The peculiarities of the Tibetan script described above complicated ATR approaches for Tibetan historical publications, particularly newspapers. Recent advances in machine learning allow us to use and train custom models for HTR—as usually applied to handwritten material such as letters or diaries—for the recognition of historical Tibetan texts.

Currently, two platforms, eScriptorium (Stokes *et al.* 2021) and Transkribus (Kahle *et al.* 2017), provide access to model training

[14] No 25 of Appendix 1 in Sawerthal 2018: 345.

[15] No 22 of Appendix 1 in Sawerthal 2018: 345.

[16] For a contemporaneous overview, see Kolmaš 1962: 638–641; Schubert 1958: 17–19.

through easy-to-use interfaces. These allow researchers unfamiliar with programming languages and computational methods to train specific models for their respective datasets. eScriptorium has higher demands for the local IT infrastructure and maintenance,[17] while Transkribus offers its platform as a service. Therefore, Transkribus is more economical for small projects such as Divergent Discourses.[18]

In Tibetan Studies, the Austrian Academy of Sciences, with its Dawn of Tibetan Buddhist Scholasticism (TibSchol) project, pioneered the development of Tibetan HTR with Transkribus. TibSchol has made public two Tibetan HTR models for Tibetan cursive, i.e. *ume* scripts.

Among the "fundamental decisions" made by the project was to start with "training a script-specific model" that later on can serve as a base model and hence significantly reduce the amount of ground truth needed for the training of other more specific models (Griffiths 2024: 45, 50). Another fundamental choice made by the TibSchol project was the decision – given the already available transcriptions (Griffiths 2024: 45) – to train the model to transcribe into Wylie, the most common romanisation system for Tibetan (Wylie 1959). Moreover, TibSchol opted for a redacted or diplomatic transcription that omits certain punctuation marks, such as the *ying-go* (*yig mgo*) or text-filling dots. The advantage of the Transkribus platform is that it

[17] Chagué and Clérice (2023) describe the technical requirements to set up eScriptorium. While it is, in principle, possible to work with a local installation on a PC without a Graphic Processing Unit (GPU), a dedicated server with a GPU is recommended for model training and multiple users. Moreover, installation, updates, and setup for the project's requirements and adjustments over the project duration will require a system administrator.

[18] Transkribus has, in recent years, established itself as a very powerful yet accessible computational tool for transcribing handwritten documents and HTR. Its flexibility made it attractive to scholars working with under-resourced and under-researched languages and scripts, such as Tibetan. Other Tibetan and Himalayan Studies research projects across Europe currently use Transkribus, most prominently the two ERC-funded projects TibSchol (Austria) led by Pascale Hugon, see TibSchol 2022 and Griffiths 2022b, and PaganTibet (France) led by Charles Ramble, see PaganTibet 2023, and more recently Law in Historic Tibet (UK) led by Fernanda Pirie, see https://www.law.ox.ac.uk/law-historic-tibet (accessed January 15, 2025).

allows for highly specific models tailored toward the specific needs and interests of any given research project.

While greatly benefitting from the experiences of TibSchol and gratefully following many of their directions, the Divergent Discourses project, aiming for a more general model, took a different approach in some respects. Most importantly, Divergent Discourses wanted to avoid working with Wylie and instead use Tibetan Unicode to avoid downstream complications. This seemed not least important since some sources used featured text in Latin script, mostly English. Moreover, we wanted to retain – as far as possible – all information from the original sources, adopting what we called a What-you-see-is-what-you-transcribe approach and decided to train a model from scratch.[19]

The Divergent Discourses project is not concerned with traditional Tibetan block prints (*dpe cha*) but with newspapers, a medium that in Tibetan areas was still emerging in the 1950s and thus was highly inconsistent in how layout principles were implemented. To deal with the complex and inconsistent, and, hence, challenging layouts, the project needed to move away from standard HTR workflows and develop a novel approach to (a) deal with the challenges posed by the Tibetan writing system, (b) handle the complex and inconsistent layouts of newspapers, and (c) enable the extraction of both text and structural information. Consequently, a four-step workflow was developed that consists of:[20]

(1) Training of HTR base model for transcribing Modern Tibetan (see section 3).

[19] We allowed one major exception to this rule by transcribing *ume* in the newspapers into *uchen* in the transcripts. The main reason behind this decision was, among others, that the great variety of *ume* scripts is not always available in Unicode, and the project wanted to avoid downstream font incompatibilities.

[20] Note: This list reflects the steps in the development of our models or, rather, in attempting to overcome challenges. With the trained models the workflow is reduced to four steps: (1) detection of structural elements, (2) detection of line polygons, and (3) HTR.

(2) Training HTR model for transcribing Tibetan Newspapers using the base model trained in step 1 (see section 4.2)
(3) Training of a Field Model (FM) for the detection of structural elements in complex newspaper layouts (see section 4.3)
(4) Training of FM for Line Polygon detection to identify text lines (see section 4.4).

3 *Training the Base model Tibetan Modern Uchen Print (TMUP)*

The stability in appearance and design and the substantial similarity of the Tibetan typefaces used – as pointed out above – led to the project’s decision to first train a robust Transkribus model for a Tibetan modern *uchen* print type, relatively standard in publications from within the PRC. This decision was inspired by the success of a related project on Uyghur newspapers, which experimented with using more general base models to train script-specific models (Barnett *et al.* 2022; Barnett & Faggionato 2022). A base model trained on a curated set of Ground Truth, i.e., accurate and verified data, accumulates and generalises knowledge about the language. We wanted to use this general *uchen* model as a base model for training more specialised models tailored for newspapers or different sets of newspapers, each with its own *uchen* typeface.

Rachael Griffiths (2024: 45–48) poignantly described the general approach to model training for Tibetan script which we generally followed:

(1) Selection of training data
(2) Annotation and Training of Layout Analysis model
(3) Adding Transcriptions and Training of HTR model

3.1 Selection of Training Data and Pre-processing

We wanted the initial model to be as robust and universal as possible so that it could be reused as a base model for more specific models trained on other sets of training data, e.g., a specific newspaper such as the *Bod ljongs nyin re'i gsar 'gyur* ("Tibet Daily"), or in other Tibetan scripts, such as various forms of *ume*. The goal was to include curated training data in the model to sufficiently represent all peculiarities found in modern Tibetan texts, including Arabic and Tibetan digits, Chinese, English and Tibetan punctuation, and unusual orthography in loan words.

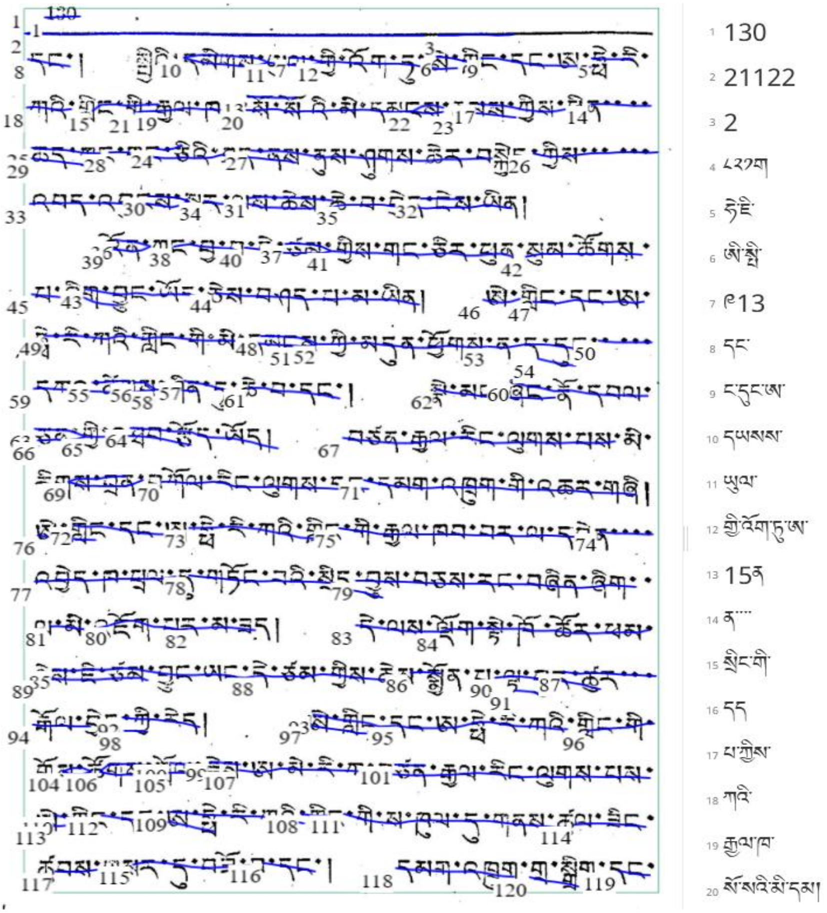
To achieve this, we decided to start with excerpts from the *Biography of Doring Paṇḍita*, an 18th-century autobiography of a Tibetan aristocrat, collated from various manuscripts and published in two volumes in 1987 in Chengdu.[21] The print of the edition is slightly clearer yet similar to Tibetan publications printed in the 1950s and 1960s. Since a digital version of the edition is available from the Buddhist Digital Resource Centre (BDRC), and Christoph Cüppers (Lumbini International Research Centre, Nepal) generously made available to us a gold-standard transcription of the text into Wylie, we anticipated saving time and cost-intensive manual transcribing work.

3.2 Layout or Baseline Model Tibetan Modern Print (TMP)

During Ground Truth transcription and experiments with HTR it became apparent that a correct layout and baseline detection is paramount for the outcome of HTR.
Three fundamental aspects must be addressed in layout recognition:

(1) The baseline is the basis for calculating line polygons. The model may miss super- or subjoined letters or vowel signs if the baseline position is incorrect.

[21] Rdo ring 1987. For more on the text and its author see Erhard 2020a; Erhard 2020b.



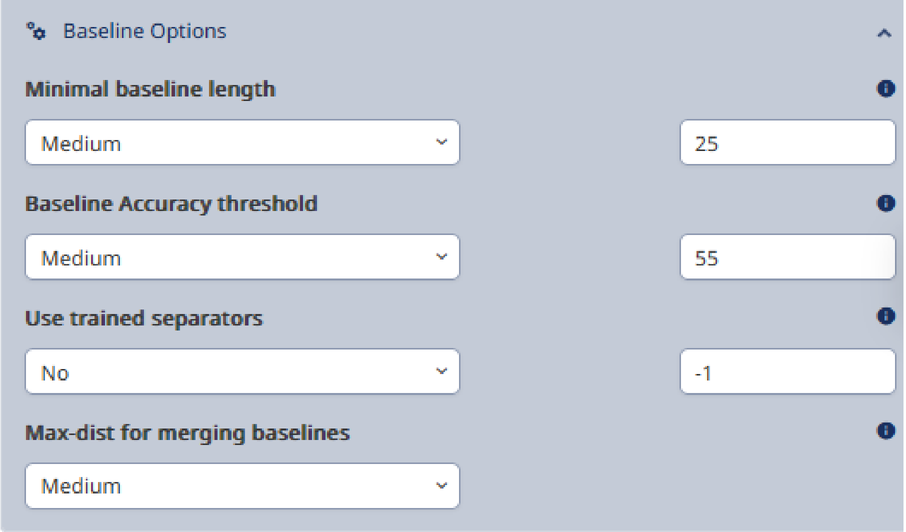
1 130
2 21122
3 2
4 ㄥ༢༧ག
5 ཧེ་ཧྲི་
6 ཨེ་སྦྲི་
7 ༡13
8 དང་
9 ང་དྭང་ཨ་
10 དཡམས་
11 ཡུལ་
12 གྱི་འོག་ཏུ་ཨ་
13 15ན
14 ན····
15 སྲིང་གི་
16 དད
17 པ་གྲིས་
18 ཀའི་
19 རྒྱལ་ཁ་
20 སོ་སའི་མི་དམ།

Figure 3 Results of HTR without previous baseline recognition

(2) In some cases, Tibetan writing has longer interspersed gaps that do not indicate a break. If the model interprets these gaps as breaks, it will introduce incorrect line breaks. In practice, it seems that the standard settings of the layout detection cause the baseline model to introduce such incorrect breaks and to draw too short baselines that miss out characters or words at the beginning and end of a line.

(3) An incorrect, e.g., too short, baseline tends to “confuse” the HTR model, resulting in incorrect transcriptions and a higher character error rate (CER).

Mitigating these issues requires a robust baseline model for Tibetan modern printed texts to be run before the actual text recognition process is started. In our case, we trained a baseline model for printed books of the second half of the 20th century in an iterative process starting with pages manually annotated during the creation of Ground Truth from excerpts from the *Biography of Doring Paṇḍita.*



Baseline Options

Minimal baseline length

Medium

25

Baseline Accuracy threshold

Medium

55

Use trained separators

No

-1

Max-dist for merging baselines

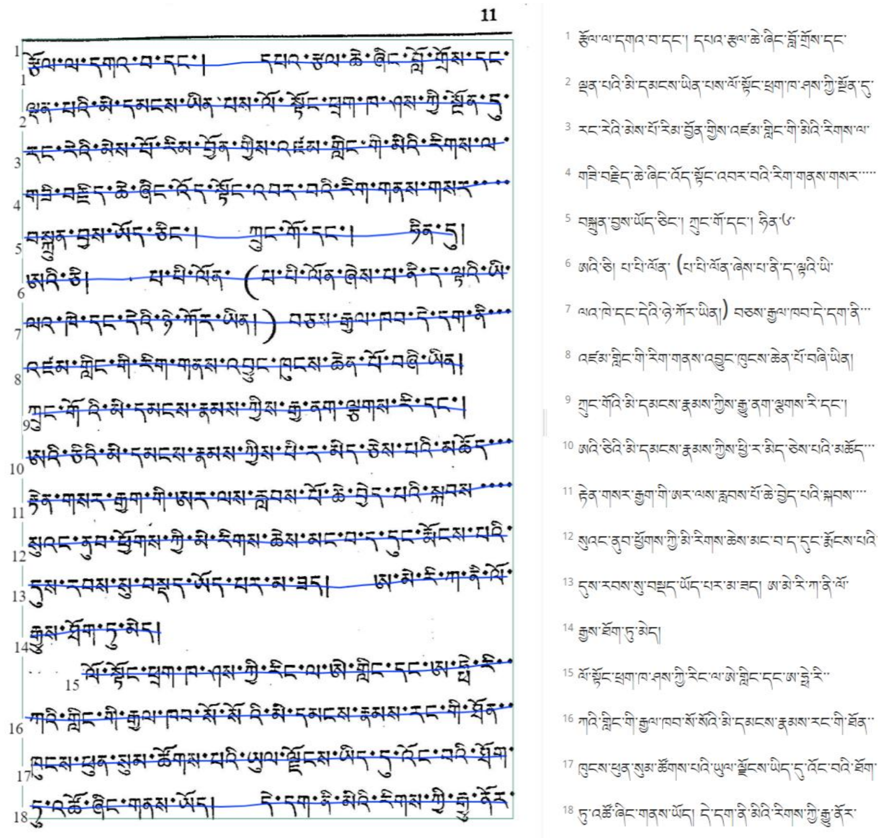
Medium

Figure 4 Baseline default options in theTranskribus advanced settings

Subsequently, the trained model was tested on unseen pages, which, after manual correction, were then added to the Ground Truth for the next iteration of the model training. With a total of 440 pages in the training set, the baseline model Tibetan Modern Print (TMP) 4.3 yielded sufficiently accurate results.[22]

Second, the layout detection settings must be adjusted to meet the specificities of Tibetan printing/writing conventions.

[22] The TMP4.3 model (Transkribus model ID 59417) is publicly available within the Transkribus platform. It was trained on curated training data from books published in the PRC between the 1950s and 1980s that include all major layout types. The training set consists of 440 pages, and the validation set consists of 37 pages. The CER measured by Transkribus is given as 3.87%.



1 རྩོལ་ལ་དགའ་བ་དང་། དཔའ་རྩལ་ཆེ་ཞིང་བློ་གྲོས་དང་

2 ལྡན་པའི་མི་དམངས་ཡིན་པས་ལོ་སྟོང་ཕྲག་ཁ་ཤས་ཀྱི་སྔོན་དུ་

3 རང་རེའི་མེས་པོ་རིམ་བྱོན་གྱིས་འཛམ་གླིང་གི་མིའི་རིགས་ལ་

4 གཞི་བརྟེན་ཆེ་ཞིང་འོད་སྟོང་འབར་བའི་རིག་གནས་གསར་·····

5 བསྐྲུན་བྱས་ཡོད་ཅིང་། ཀྲུང་གོ་དང་། ཧིན་༤་

6 ཨའི་ཅི། པ་པི་ལོན་ (པ་པི་ལོན་ཞེས་པ་ནི་ད་ལྟའི་ཡི་

7 ལའ་ཁེ་དང་དེའི་ཉེ་གོར་ཡིན།) བཅས་རྒྱལ་ཁབ་དེ་དག་ནི···

8 འཛམ་གླིང་གི་རིག་གནས་འབྱུང་ཁུངས་ཆེན་པོ་བཞི་ཡིན།

9 ཀྲུང་གོའི་མི་དམངས་རྣམས་ཀྱིས་རྒྱ་ནག་ལྷུགས་རི་དང་།

10 ཨའི་ཅིའི་མི་དམངས་རྣམས་ཀྱིས་ཕྱི་ར་མིད་ཅེས་པའི་མཆོད···

11 ཧེན་གསར་རྒྱག་གི་ཨར་ལས་རླབས་པོ་ཆེ་བྱེད་པའི་སྐབས····

12 སྤུའང་ནུབ་ཕྱོགས་ཀྱི་མི་རིགས་ཆེས་མང་བ་ད་དུང་རྨོངས་པའི

13 དུས་རབས་སུ་བསྡད་ཡོད་པར་མ་ཟད། ཨ་མེ་རི་ཀ་ནི་ལོ་

14 རྒྱུས་ཐོག་ཏུ་མེད།

15 ལོ་སྟོང་ཕྲག་ཁ་ཤས་ཀྱི་རིང་ལ་ཨེ་གླིང་དང་ཨ་ཧྥི་རི··

16 ཀའི་གླིང་གི་རྒྱལ་ཁབ་སོ་སོའི་མི་དམངས་རྣམས་རང་གི་ཐོན··

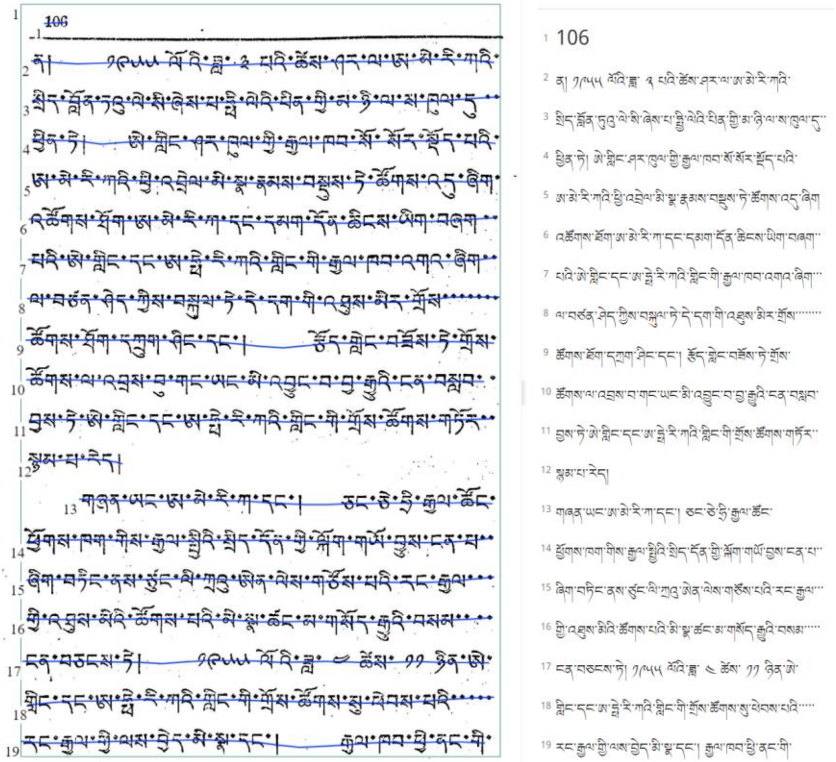
17 ཁུངས་ལྡན་སུམ་ཚོགས་པའི་ཡུལ་ལྗོངས་ཡིད་དུ་འོང་བའི་ཐོག་

18 ཏུ་འཚོ་ཞིང་གནས་ཡོད། དེ་དག་ནི་མིའི་རིགས་ཀྱི་རྒྱུ་ནོར་

Figure 5 Results of baseline recognition TMP4.3 with standard settings and subsequent HTR

The examples in Figures 5 and 6 illustrate how baseline recognition with different settings affects the accuracy of subsequent HTR. The standard settings for baseline recognition in Transkribus (see Fig. 4 above) yield excellent results for our corpus of Tibetan printed publications from the 1950s to 1980s. Only the page number is missed by the model. Yet, missing out on page numbers or orphaned syllables potentially poses a serious problem for text extraction. Reducing the Minimum Baseline Length in the advanced settings from Medium (25) to Low (10) enables the model to catch the page number in line 1 (see Fig. 6).

Besides being usable for the transcription of modern Tibetan print publications with only minor manual correction, we assume that future models by the project will overcome the current shortcomings by including additional training data, which can now be quickly produced using the current Tibetan Uchen Print (TMP) 4.4 model.



106
ནེ། ༡༩༥༥ ལོའི་ཟླ་ ༤ པའི་ཚེས་ཤར་ལ་ཨ་མེ་རི་ཀའི་
སྲིད་བློན་ཏུའུ་ལེ་སི་ཞེས་པ་སྤྱི་ལེའི་ཕིན་གྱི་མ་ཉི་ལ་ས་ཁུལ་དུ་་
ཕྱིན་ཏེ། ཨེ་གླིང་ཤར་ཁུལ་གྱི་རྒྱལ་ཁབ་སོ་སོར་སྡོད་པའི་
ཨ་མེ་རི་ཀའི་ཕྱི་འབྲེལ་མི་སྣ་རྣམས་བསྡུས་ཏེ་ཚོགས་འདུ་ཞིག
འཚོགས་ཐོག་ཨ་མེ་རི་ཀ་དང་དམག་དོན་ཆིངས་ཡིག་བཞག་་
པའི་ཨེ་གླིང་དང་ཨ་ཧྲི་རི་ཀའི་གླིང་གི་རྒྱལ་ཁབ་འགའ་ཞིག་་་
ལ་བཙན་ཤེད་ཀྱིས་བསྐུལ་ཏེ་དེ་དག་གི་འཛུལ་མིར་གྲོས་·······
ཚོགས་ཐོག་དཀྲུག་ཤིང་དང་། རྫོད་གླེང་བཟོས་ཏེ་གྲོས་
ཚོགས་ལ་འཐུས་པ་གང་ཡང་མི་འབྱུང་བ་བྱ་རྒྱུའི་ངན་བསླབ་
བྱས་ཏེ་ཨེ་གླིང་དང་ཨ་ཧྲི་རི་ཀའི་གླིང་གི་གྲོས་ཚོགས་གཏོར་་
སློམ་པ་རེད།
གཞན་ཡང་ཨ་མེ་རི་ཀ་དང་། ཅང་ཅེ་ཧྲི་རྒྱལ་ཚོང་
ཕྱོགས་ཁག་གིས་རྒྱལ་སྤྱིའི་སྲིད་དོན་གྱི་སྐོག་གཡོ་བྱས་ངན་པ་་
ཞིག་བཏིང་ནས་ཙུང་ལི་ཀྲའུ་ཨེན་ལེས་གཙོས་པའི་རང་རྒྱལ་་
གྱི་འཛུལ་མིའི་ཚོགས་པའི་མི་སྣ་ཚང་མ་གསོད་རྒྱུའི་བསམ·····
ངན་བཅངས་ཏེ། ༡༩༥༥ ལོའི་ཟླ་ ༤ ཚེས་ ༡༡ ཉིན་ཨེ་
གླིང་དང་ཨ་ཧྲི་རི་ཀའི་གླིང་གི་གྲོས་ཚོགས་སུ་ཐེབས་པའི·····
རང་རྒྱལ་གྱི་ལས་བྱེད་མི་སྣ་དང་། རྒྱལ་ཁབ་ཕྱི་ནང་གི་

Figure 6 Transkribus screen shot showing perfect results for baseline detection and subsequent HTR with TMUP 0.1

3.3 *Training of a Tibetan HTR Model*

3.3.1 *Step 1: Iterative Training*

The training of a Tibetan HTR model for Tibetan has been described recently (Griffiths 2024) and will not be repeated here in detail. At the time of writing, the Transkribus platform has almost fully transitioned from the Transkribus Expert Client (and accompanying Transkribus

Lite) to the Transkribus web app, and many of the settings described by Griffiths seem to be handled automatically in the app by now.[23]

During our training of the base model Tibetan Modern U-chen Print (TMUP), to check the accuracy of the results and fine-tune the model we iterated through a series of training runs. After each training run, the results were evaluated, errors corrected, and deficiencies identified. The corrected material was then added to the training data.

As detailed above, we started with 63 pages from the *Biography of Doring Paṇḍita,* gradually adding more pages from books mostly published in the PRC in the 1950s, including Liu Shaoqi's (1898–1969) *Marxism-Leninism is Victorious in China* (Li'u hra'o chis 劉少奇 1959), Mao Zedong's (1893–1976) *Treaty on New Democracy* (Ma'o tse tung 毛澤東 1952), etc., but also Gendün Choephel's (1903–1950) *Guidebook to India* (Dge 'dun chos 'phel 1968).[24] The titles were selected because they reflected the Diverge Discourses' time frame and thematic focus. More importantly, they contained a wide variety of typographical signs, punctuation, and orthographies particular to the 1950s and 1960s.

3.3.2 Step 2: Training of Tibetan Modern U-Chen Print (TMUP) 0.1

The next and final step required us to manually evaluate the training results and test the model on unseen data. This process involved several intermediate steps to identify deficiencies in the model.

The material of the project's research period is challenging as it is characterised by rapid social, technological, and, subsequently, linguistic development. Social and political change made it necessary to adopt new terminologies, which often came to Tibetan as loanwords either from Chinese, English, or Hindi. Some of these new words,

[23] With the transition to the Transkribus web app, most of the settings for model training have become inaccessible for the user of the app. For example, the selection of a de-warping method or the batch size in the advanced settings for model training cannot be changed in the web app.

[24] For an inspection of the full set of training data, see Erhard *et al.* 2024.

particularly toponyms and anthroponyms, made including new sounds in the Tibetan language necessary.

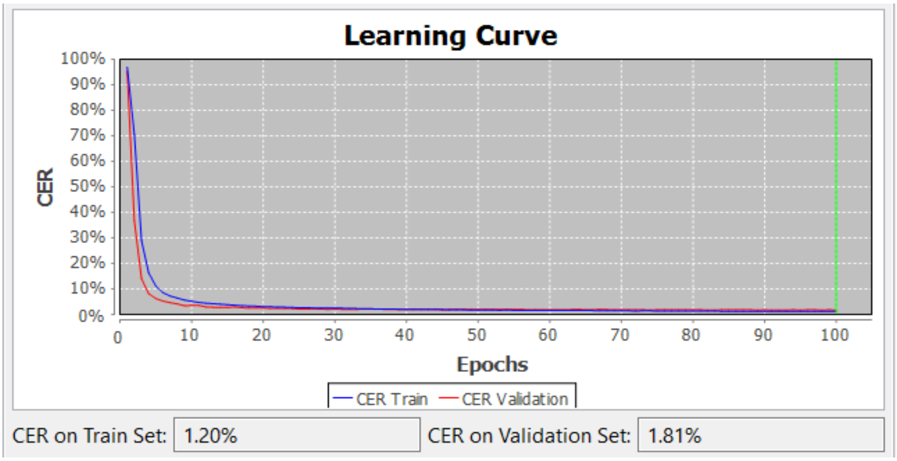
A particular difficulty was including enough data containing the wide variety of punctuation marks, digits, and signs stemming from Tibetan, Chinese and English writing conventions. Also, we assumed that the rare—but still in occasional usage—long stacks of e.g., Sanskrit terms, but also the in modern Tibetan widespread use of unusual orthography for foreign names or loanwords, such as the stack *hpha* ཧྥ (rendering the labial fricative "f") in *hpha ran zi* ཧྥ་རན་ཛི (France), but also unusual orthography in Chinese names such as Le'u Hro'o chi ལེའུ་ཧྲོའོ་ཆི (Liu Shaoqi 劉少奇), or the Chinese appellation *hru'u ci* ཧྲུའུ་ཅི (*shuji* 书记 "secretary"), etc. need their fair representation in the training data. Finally, we discovered that in publications originating from Tibet and China, especially in the 1950s, a wide range of punctuation marks, including various quotation marks and brackets used in traditional Chinese, were used, while in the *Tibet Mirror*, the leading Tibetan newspaper published on the subcontinent, punctuation marks were often borrowed from the English usage (see the Appendix of transcribing conventions).

To overcome these, we added specifically curated material, such as material found in the bibliographic information of book publications, that contained missing letters, characters, or specific signs to the training data.

That way, we produced more transcriptions, which could quickly be corrected for Ground Truth. Gradually, we enlarged the training data set to 522 pages from twenty different sources published in the PRC between the 1950s and 1980s, as well as a few exceptional pages published in India, to ensure that all special characters, particularly Tibetan and Arabic numbers, are contained in the training set.

We trained the final model without using an existing model as the base model to ward off unpredicted behaviour or unwanted interferences. The training set consists of 470 pages; the validation set consists of 52 (10%) automatically selected pages.[25]

The resulting model Tibetan Modern U-chen Print 0.1 (TMUP 0.1) validated with a CER of 1.81% and is the first Transkribus HTR model for printed Tibetan language publications in *uchen* script. As the learning curve in Figure 7 suggests, the model is already close to overfitting, which was avoided by automatically stopping the training at 100 epochs.[26]



CER on Train Set: 1.20% CER on Validation Set: 1.81%

Figure 7: Tibetan Modern U-chen Print (TMUP) 0.1 Learning Curve

Despite the model’s good CER, it still had difficulties in unseen text with subscribed letters, especially subscribed signs for the vowel “u” ◌ུ (*zhabs skyu*) and subscribed consonants “y” ◌ྱ (*ya brtags*) or “r” ◌ྲ (*ra brtags*) in long stacks.[27]

[25] See the full training set in Erhard *et al.* 2024. To reserve 10% of the material as a validation set is the standard option in Transkribus, following the general recommendation of 10-fold cross-validation in model training.

[26] In Transkribus, the predefined settings automatically stop (“early stopping”) HTR training if the learning curve does not improve after 20 epochs. This generally prevents overfitting of the model to the training data.

[27] This is a result of the automatic creation of line polygons, i.e., the text fields drawn around the whole line of text including ascenders and descenders, from the baselines. For a more detailed discussion of this problem, see section 4.4 below.

Nevertheless, TMUP 0.1 now transcribes modern Tibetan text with only minor errors, which usually occur only with less frequent syllables, such as phonetic transliterations of Chinese or Western names and terms such as *cang ce hyi* ཅང་ཅེ་ཧྱི (for *cang ci hri* ཅང་ཅི་ཧྲི) or *le ze thun ho zi gung zi* ལེ་ཟེ་ཐུན་ཧོ་ཟི་གུང་ཟི (for *we zi than ho zi gun zi* ཝེ་ཟི་ཐུན་ཧོ་ཟི་གུང་ཟི), occasionally confusing the very similar vowel signs for *i* ◌ི and *e* ◌ེ or also very similar syllables *rnga* རྔ and *ja* ཇ. Another source of misreading stems from unclear print, such as *bcongs* བཅོངས (for *btsongs* བཙོངས) or *klog* ཀློག (for *glog* གློག).

Interestingly, when testing the model with Tibetan publications from India, the model struggled with ordinary text. Indian publications feature a similar yet slightly different typeface as described above in section 2.1. Although the differences may seem irrelevant to a human reader, the difficulties of the model dealing with Indian publications indicate it was overfitting to the material from the PRC thus introducing a strong bias towards PRC typefaces. Moreover, this indicates that curating similar data in the training set reduces the model's generalising abilities.

4 *Tibetan Newspapers: Text Recognition and Information Extraction*

The above-described procedures yielded satisfying results for publications with simple layouts, such as Western-style books and homogenous typefaces. However, when it came to Tibetan newspapers, with their varied typefaces and complex and often inconsistent layouts, i.e., steps 2-4 in the workflow outlined above, our custom-trained baseline and HTR models were insufficient to capture the material's complexities and produce acceptable results.

4.1 *Newsprint as an Exceptional Case*

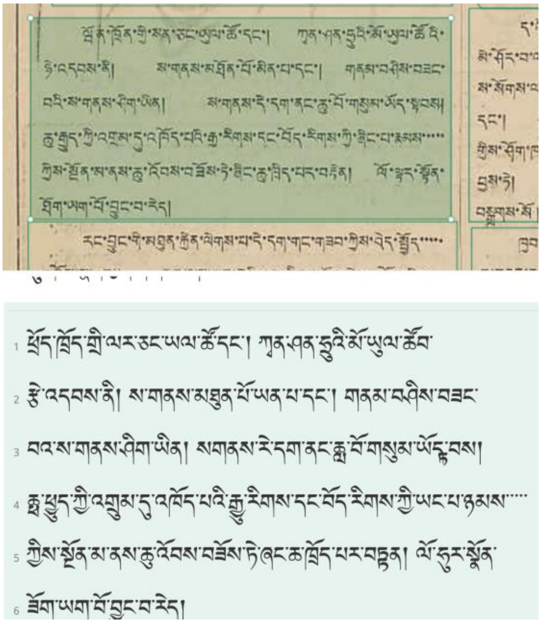
The Divergent Discourses newspaper corpus presented us with two unrelated problems. First, as mentioned above, our HTR base model could not handle the various scripts and font types used in Tibetan newsprint. Second, the newspapers' complex and varied column layout cannot be handled with a standard baseline model. Since Transkribus's computer vision algorithm reads pages from top-left to bottom-right, it struggles with newspaper columns. It mistakenly jumps from the first line in the first column to the first line in the second column and so forth, reading lines of the same height in different columns as a single line. Subsequent text recognition then produces an e-text with a jumbled reading order.

4.2 Handwritten Text Recognition for Tibetan Newspapers

The project manually transcribed 40 pages of Ground Truth for each of the eleven core newspapers to deal with different scripts and fonts in Tibetan newspapers.[28] We tested two approaches: (1) Train many newspaper-specific models, and (2) Train one model that can handle all newspapers (One4All).

(1) With enough ground truth for some newspapers, we trained a specific model using TMUP 0.1 as a base model. The resulting models had CERs between 1.9% and 7.2%, which seemed acceptable given the small training set (30–40 pages per newspaper). However, testing the models on unseen data quickly showed that they were performing poorly. Fig. 8 shows an example page from Minjiang News transcribed with a model (minjiang v01 ID 59357) trained on 37 pages of the same newspaper. The short paragraph clearly shows that the model fails to transcribe the text correctly. Consequently, more Ground Truth is need-ed to achieve satisfactory results.

[28] While the Divergent Discourses Corpus contains 17 different newspaper titles, not all titles are available in sufficient quantities (e.g. only one issue or four pages of Gyantse News GTN) or were mostly in the Chinese language (e.g. the newspaper of the Central Institute for Nationalities ZMX), for pragmatic reasons we limited the training data to a core of eleven newspapers. For a description of the Divergent Discourses Corpus, see Erhard 2025 in this special issue.



སྤྲོད་ཁྲོད་གྱི་ལར་ཅང་ཡལ་ཆོ་དང་། ཀུན་ཤན་ཧྲུའི་མོ་ཡུལ་ཆོབ་
རྩེ་འདབས་ནི། ས་གནས་མཐུན་པོ་ཡན་པ་དང་། གནམ་བཞིས་བཟང་
བའ་ས་གནས་ཞིག་ཡིན། སགནས་རི་དག་ནང་ཀླ་བོ་གསུམ་ཡོད་རྐྱབས།
ཆུ་ཕྱུད་ཀྱི་འགྲུལ་དུ་འགོད་པའི་རྒྱུ་རིགས་དང་བོད་རིགས་ཀྱི་ཡང་པ་ཉམས·····
ཀྱིས་སྤོན་མ་ནས་ཆུ་འོབས་བཟོས་ཧེ་ཞང་ཆ་ཁྲོད་པར་བརྟན། ལོ་ཧུར་སྔོན་
ཐོག་ཡག་བོ་བྱུང་བ་རེད།

Figure 8 Perfomance of model minjiang v01 (ID 59357) with a CER of 0.5 on unseen text

(2) Combining all available training data for a One4All model exposes the PyLaia algorithm behind Transkribus with a greater variety of scripts, fonts and layouts. While there might be some risk of "confusing" the algorithm, it also increases the model's "knowledge" about scripts and fonts. The resulting model TibNews-One4All 0.1, trained with TMUP 0.1 as a base model on 269 pages (42,503 words[29]), initially showed a CER of 3.3%. The latest version TibNewsOne4All 0.2, trained with TMUP 0.1 as a base model on 500 pages (92,423 words), has a CER of 2.52%.[30]

While the CERs are similar, the TibNews-One4All model performed much better on unseen newspaper material and unrelated material from an earlier period.[31] The model's good performance on unseen material indicates that more variety in the training set enhances the model's ability to generalise.

[29] The term "words" is inherited from Transkribus but in the context of the Tibetan language confusing. Transkribus probably simply takes every token separated by whitespace as a word. Therefore, most likely it is Tibetan syllables that Transkribus interprets as words.

[30] The TibNewsOne4All 0.2 (ID 169581) is publicly accessible in Transkribus (https://www.transkribus.org/model/tibnewsone4all, accessed January 14, 2025)

[31] Daniel Wojahn, in the context of the project *Law in Historic Tibet* (Oxford), tested the model on Tibetan legal texts and reported very good results.

4.3 *Identifying and Classifying Structural Elements with Transkribus's Field Models*

A more efficient layout analysis is necessary to solve the problem of complex column layout. In late 2023, Transkribus slowly started introducing models for advanced layout analysis and information extraction, including trainable field models (FM).[32] These allow the detection of different text regions, such as columns, paragraphs, etc., on a single page and restrict baseline detection and subsequent text recognition to these regions.

During ground truth transcription for training HTR models, the Diverge project manually annotated columns and other structural elements in 500 Tibetan-language newspaper pages. Since the field models can be trained to identify layout elements, we refined the annotation with the following structural elements: page numbers, headers, newspaper titles, headings, captions, paragraphs, marginalia, and other generic elements.[33] Additionally, we experimented with labelling text in English and vertical and horizontal Chinese. These labels describe the main structural elements in the newspapers and constitute important information we would like to retrieve automatically.[34]

That way, our field model TibNewsTR 0.6.5 (ID 232709), trained on 609 pages, could recognise and classify the differing text regions. The model has a mean average precision (mAP) of 47.96%. Although a mAP of less than 50% indicates that the model's classification abilities are still relatively low, the accuracy of identifying text regions and, consequently, handling complex column layouts is much higher.[35]

[32] Field Models are only available with a Transkribus subscription, starting with a scholar plan. Moreover, at the time of writing, FMs cannot be used via the API; this is likely to change over the coming months.

[33] For a comprehensive tag list, see the appendix, section 5.1.

[34] While Transkribus can identify and label these structural elements, and store the information in the output PAGE XML, the actual retrieval must be done in post-processing.

[35] Transkribus provides no evaluation score for simple text region detection.

For the subsequent text recognition, baseline models can be set up to split lines at the text region borders, allowing for handling complex column layouts.

4.4 Handling Different Font Sizes with Line Polygon Models

Although we could now transcribe Tibetan newspapers with complex layouts, the HTR model struggled to correctly transcribe several features of our sources, such as longer stacks, headlines in larger font, or vertical text.

4.4.1 The Problem

In the early 1950s, many newspapers featured text written in both horizontal and vertical Chinese. To be able to correctly identify all text, we needed a model that could handle left-to-right (LTR) text as well as Chinese text written from top to bottom and right-to-left (RTL).

A second problem is directly affecting Tibetan text recognition. As mentioned in section 3.3.2, our HTR models struggled with longer stacks, particularly with subscribed letters and vowels; they also failed to transcribe large print newspaper titles and headlines. Interestingly, the automatically calculated line polygons were often too small to include longer stacks and often covered only the core area of large print headlines or newspaper titles (Fig. 9). This can be explained through the automatic calculation of line polygons, i.e., the outline of the whole text line, from baselines, which assumes a text of homogenous font size. Consequently, layout analysis with custom-trained baseline models effectively processes standard text, i.e. text of the same orientation and size on which the model had been trained. However, it struggles with text in different orientations and sizes and consequently, the results of HTR are unsatisfactory for these text regions (Fig. 10).

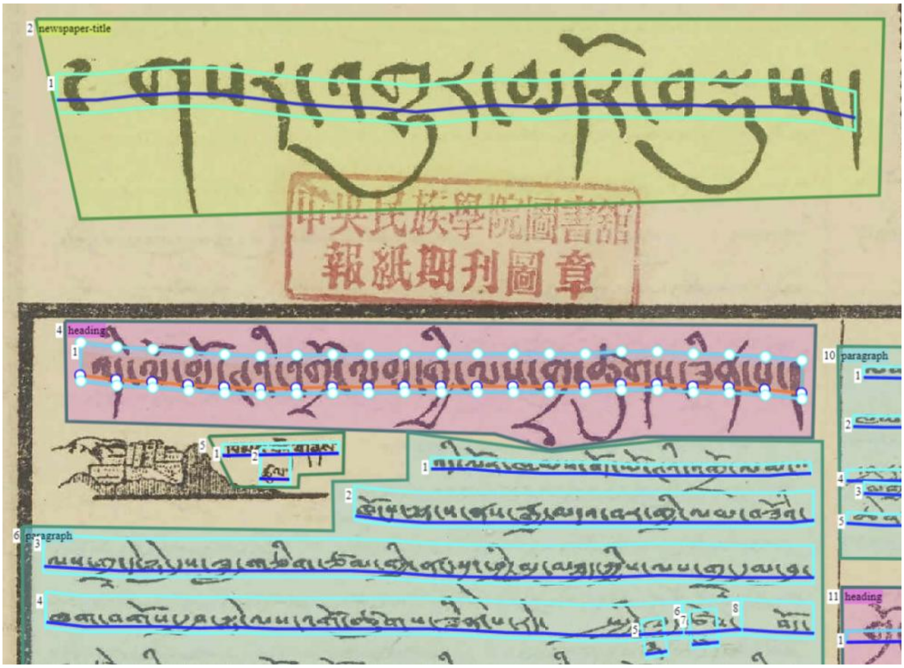
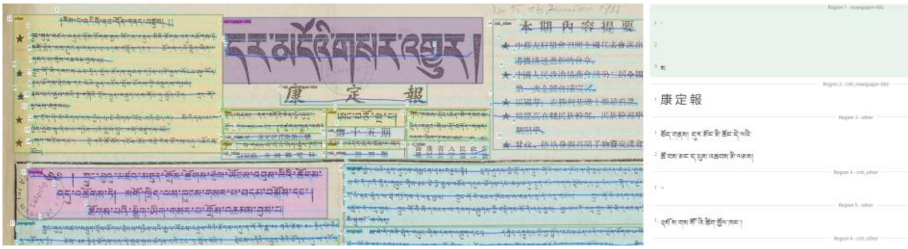


Figure 9 Line polygons (turquoise), automatically calculated from baselines (blue)

4.4.2 Dedicated Field Models for Line Polygon Recognition

With the introduction of FM, an alternative way of line recognition was introduced to Transkribus. As outlined above, FM can be trained to identify text regions and classify them as structural elements. However, FM can also be trained on manually annotated line polygons. Provided enough training data, line polygon FM can then detect exact line polygons, including the upper (ascenders) and lower (descenders) reach of longer strokes, stacks or vowels. With the new field models, more traditional layout/baseline detection becomes obsolete, and subsequent HTR "searches" for all text within each line polygon, theoretically allowing for the recognition of vertical text.[36]

[36] This approach was suggested by the Transkribus team following a roundtable on Transkribus for Asian Languages at TUC24 organised by Rachael Griffith and



Region 1 - newspaper-title

1

2

3 ས

Region 2 - chi_newspaper-title

1 康定報

Region 3 - other

1 ཚོད་གནས། དུས་ཚོང་གི་ཚོང་དེ་ལའི

2 སྐྱེ་བས་ཟང་ད་བྱུས་འཚམས་ཀྱི་ལཀ་ས།

Region 4 - chi_other

1 ..

Region 5 - other

1 དགོས་གས་གོ་རི་ཚིག་རྩོམ་ཁང་།

Region 6 - chi_other

Figure 10 HTR after baseline recognition: Region 1, the newspaper-title, has been incorrectly recognised as three lines with narrow line polygons. Consequently, the HTR model has failed to transcribe this correctly.

To solve the issues with RTL Chinese and longer stacks, we trained a FM on line polygons on 121 manually annotated pages. The resulting model FM TibNewsLines 0.2.2 (ID 169109) showed a mAP of 50.59%. The relatively low mAP reflects low confidence scores for vertical Chinese due to a lack of training data.

Moving away from the standard HTR workflow – baseline recognition followed by HTR – to a more complex three-step workflow starting with running the FM TibNewsTR 0.6.5 for text region recognition and classification, followed by the FM TibNewsLines 0.2.2 to detect line polygons, and finally, the HTR model TibNewsOne4All 0.2 drastically improved the results (Fig. 11).

With more Ground Truth gradually becoming available, we expect future models of the project to be able to handle text in other languages/scripts, particularly vertical Chinese.

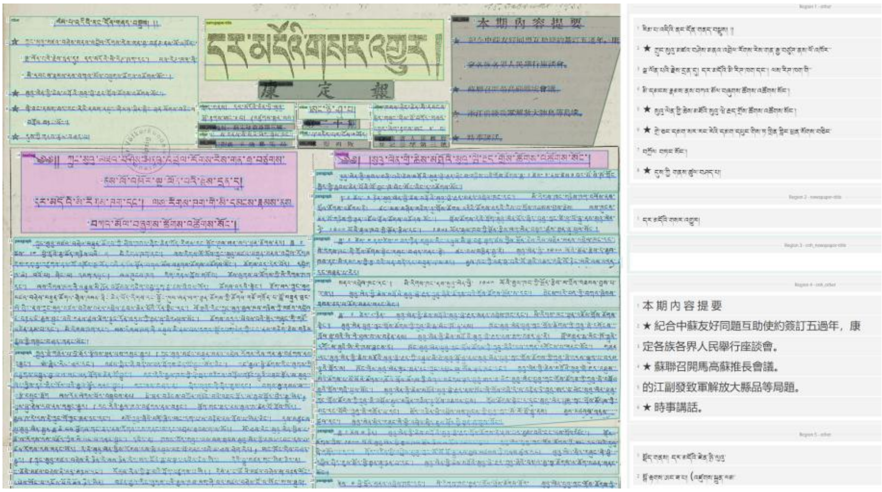
5 *Conclusion*

Platforms like Transkribus offer HTR, an efficient and affordable solution for smaller research projects like the Divergent Discourses project.

Franz Xaver Erhard (Griffiths *et al.* 2024). At the time of writing, our HTR models could not sufficiently transcribe Chinese to evaluate the line polygon FM's performance on vertical text.

No out-of-the-box solution to Tibetan automatic text recognition is likely to become available soon, given the vastness of Tibetan literature and the variation in printing technologies, scripts and types.

The described workflow of the Divergent Discourses project demonstrates three crucial points for automated text recognition and corresponding model training. First, accurate layout detection is fundamental to the text recognition process. Depending on the source material, Transkribus field models can solve (1) the problem posed by complex layouts, such as in newspapers, and (2) problems of ascenders and descenders frequent in the long stacks of Tibetan writing.



Region 1 - other

ཞིབ་པ་འདིའི་ནང་དོན་གནད་བསྡུས། །

★ [illegible]

[illegible]

[illegible]

★ [illegible]

★ [illegible]

[illegible]

★ [illegible]

Region 2 - newspaper-title

དར་མདོའི་གསར་འགྱུར།

Region 3 - zh_newspaper-title

Region 4 - zh_other

本期内容提要

★紀合中蘇友好同題互助使約簽訂五週年，康

定各族各界人民舉行座談會。

★蘇聯召開馬高蘇推長會議。

的江副發致軍解放大縣品等局題。

★時事講話。

Region 5 - other

[illegible]

[illegible]

Figure 11 HTR after detection of line polygons with FM TibNewsLines 0.2.2:

Second, using a base model for HTR model training speeds up the initial training process. However, it should be noted that this advantage only holds for initial training with little training data. Once more training data is available, similar CERs can be achieved with and without using a base model in training, as was shown with TMUP 0.1.

Third, highly specialised HTR models, e.g., trained on one scribal hand or, in our case, one particular newspaper, tend to perform poorly on unseen texts. Homogenous training data thus causes the model to overfit to a specific style or type of script. Conversely, models trained on a broad range of training data, including different scribal hands, font styles or scripts, have better generalisation capabilities and perform better in a broader range of sources.

Individual Tibetan digitisation projects develop their own very specialised models with varying approaches to transcribing Tibetan sources, resulting in limited reusability of the Ground Truth and models produced. Consequently, to train stronger, more capable HTR models, digitisation projects should follow a similar standard in their Ground Truth transcriptions to make training data sets more transparent and compatible. With the publication of the Divergent Discourses' transcribing conventions, we hope to provide an incentive and a starting point for the development of widely reusable HTR models for Tibetan.

Bibliography

Barnett, Robert and Christian Faggionato

"HKBUproject. historical-uyghur-chinese-corpus," *Zenodo,* 2022. doi:10.5281/ZENODO.6513855

Barnett, Robert, Jessica Yeung, Ahmet Hojam Pekiniy, Rune Steenberg Reyhe, Merhaba Eli, and Christian Faggionato

"A Resource for the Study of Translation into Uyghur by Modern Chinese Governments." *In* M. Schatz (ed.) *Multiethnic Societies of Central Asia and Siberia Represented in Indigenous Oral and Written Literature: The Role of Private Collections and Libraries*, Göttingen: Universitätsverlag,. 2022, pp. 11–27. doi:10.17875/gup2022-2054.

Bradburne, James M.

Die schwarze Stadt an der Seidenstraße: Buddhistische Kunst aus Khara Khoto (10. - 13. Jh.). Mailand: Electa, 1993.

Cabezón, José Ignacio and Roger R. Jackson

"Editors' Introduction." *In* José Ignacio Cabezón and Roger R. Jackson (eds.) *Tibetan literature: Studies in genre*. Studies in Indo-Tibetan Buddhism. Ithaca: Snow Lion, 1996, pp. 11–37.

Chagué, Alix and Thibault Clérice

"017 - Deploying eScriptorium online: notes on CREMMA's server specifications," 2023. Available online https://inria.hal.science/hal-04362085v1 (accessed January 14, 2025).

Dge 'dun chos 'phel

Rgya gar gyi gnas chen khag la bgrod pa'i lam yig [Guidebook to India's Sacred Sites]. Gsung rab bces btus dpar khang, 1968.

Erhard, Franz Xaver

“The Divergent Discourses Corpus: A Digital Collection of Early Tibetan Newspapers of the 1950s and 1960s,” *Revue d’Etudes Tibétaines* (73), 2025, pp. 44–80.

"Doring Tenzin Peljor." *Treasury of Lives*, 2020a. Available online https://treasuryoflives.org/biographies/view/Doring-Tenzin-Peljor/5306 (accessed February 09, 2021).

"Genealogy, Autobiography, Memoir. The Secular Life Narrative of Doring Tenzin Penjor." *In* Franz Xaver Erhard and Lucia Galli (eds.) “The Selfless Ego II: Conjuring Tibetan Lives,” Special issue, *Life Writing* 17 (3), 2020b, pp 327–45. doi:10.1080/14484528.2020.1737496.

"Media and Printing in Tibet since 1950. A Preliminary Survey of Tibetan Language Journals and Magazines." *In* Pavel Grokhovskiy (ed.) *Modernizing the Tibetan Literary Tradition*, Saint Petersburg: St Petersburg Univ Press, 2018, pp. 110–34.

Erhard, Franz Xaver and Haoran Hou

"The *Melong* in Context. A Survey of the Earliest Tibetan Language Newspapers 1904–1960." *In* Françoise Wang-Toutain and Marie Preziosi (eds.) *Cahiers du Mirror*. Paris: Collège de France, 2018, pp. 1–40.

Erhard, Franz Xaver, Xiaoying 笑影; Robert Barnett, and Nathan W. Hill

“Tibetan Modern U-chen Print (TMUP) 0.1: Training Data for a Transkribus HTR Model for Modern Tibetan Printed Texts. [data set],” *Fachinformationsdienst (FID) Asi*en, 2024. doi:10.48796/20240313-000.

Fader, H. Louis

Called from Obscurity. The Life and Times of a True Son of Tibet, God's Humble Servant from Poo, Gergan Dorje Tharchin: Vol. II. Kalimpong: Tibet Mirror Press, 2004.

Called from Obscurity. The Life and Times of a True Son of Tibet, God's Humble Servant from Poo, Gergan Dorje Tharchin: Vol. III. Kalimpong: Tibet Mirror Press, 2009.

Griffiths, Rachael

"Handwritten text recognition (HTR) for Tibetan Manuscripts in Cursive Script." *Revue d'Etudes Tibétaines* (72), 2024, pp. 43–51. Available online at https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_72_03.pdf (accessed January 24, 2025).

"Transkribus in Practice. Abbreviations.".*The Digital Orientalist*, 2022a. Available online https://digitalorientalist.com/2022/11/01/transkribus-in-practice-abbreviations/ (accessed January 02, 2025).

"Transkribus in Practice. Improving CER." *The Digital Orientalist*, 2022b. Available online https://digitalorientalist.com/2022/10/25/transkribus-in-practice-improving-cer/ (accessed January 14, 2025).

Griffiths, Rachael, Franz Xaver Erhard, James H. Morris, Alexander O'Neill, Li Shihua, and Nicole Merkel-Hilf

"Round Table: Transkribus for Asian Languages #TUC24 – YouTube," 2024. Available online at https://www.youtube.com/watch?v=-74AQDFaTyE (accessed December 20, 2024).

Hartley, Lauran R.

2003. "Contextually Speaking. Tibetan Literary Discourse and Social Change in the People's Republic of China (1980-2000)." Diss., Department of Central Eurasian Studies, Indiana University.

Kahle, Philip, Sebastian Colutto, Günter Hackl and Günter Mühlberger
"Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents." In *14th IAPR International Conference on Document Analysis and Recognition*. Los Alamitos: IEEE Computer Society, 2017, pp. 19–24. doi:10.1109/ICDAR.2017.307.

Kolmaš, Josef
"Tibetan Literature in China," *Archív Orientální* 30, 1962, pp. 638–644.

Kyogoku, Yuki, Franz Xaver Erhard, Robert Barnett, and Nathan W. Hill
"TibNorm - Normaliser for Tibetan (Version v1)," *Zenodo*, 2024, doi: 10.5281/zenodo.10815272.

.

Li'u hra'o chis 劉少奇
Krung gor mar khe si dang le nyin ring lugs rnams par rgyal ba [Long live Marxism and Leninism in China]. Pe cin: Mi rigs dpe skrun khang, 1959.

Luo, Queenie and Leonard W. J. van der Kuijp
"Norbu Ketaka: Auto-Correcting BDRC's E-Text Corpora Using Natural Language Processing and Computer Vision Methods," *Revue d'Etudes Tibétaines* (72), 2024, pp. 26-42. Available online at https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_72_02.pdf (accessed January 20, 2025).

Ma'o tse tung 毛澤東
Dmangs gtso'i ring lugs gsar pa'i bstan bcos. [Treatise on New Democracy]. Pe cing: Krung dbyang mi dmangs srid gzhung mi rigs don byed u yon lhan khang, 1952.

Moskaleva, Natalia N. and Pavel L. Grokhovskiy
"*The Tibet Mirror* Vol. I, No 1. Translation and Transliteration." *In* Françoise Wang-Toutain and Marie Preziosi (eds.) *Cahiers du Mirror*. Paris: Collège de France, 2018, pp. 147–168.

Nockels, Joseph, Paul Gooding, and Melissa Terras

"The implications of handwritten text recognition for accessing the past at scale," *Journal of Documentation* 80 (7), 2024, pp. 148–67. doi: 10.1108/JD-09-2023-0183.

PaganTibet

"Reconstructing the Tibetan Pagan Religion," 2023. Available online https://www.crcao.fr/recherche/pagantibet-documenter-la-premiere-reconstruction-de-pratiques-prebouddhiques-au-tibet/?lang=en (accessed January 14, 2025).

Pistorius, Kristin

"Die *Bod yig phal skad kyi gsar 'gyur*. Sprachrohr der frühen Chinesischen Republik." Masterarbeit, Institut für Indologie und Zentralasienwissenschaften, Universität Leipzig, 2019. Available online at https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa2-933730 (accessed January 14, 2025).

Rdo ring Bstan 'dzin dpal 'byor (b. 1760)

1987. *Rdo ring paṇḍi ta'i rnam thar.* [The Biography of Doring Paṇḍita]. 2 vols. Khren tu'u: Si khron mi rigs dpe skrun khang, 1987. BDRC: W1 PD96348.

Rowinski, Zach and Kurt Keutzer

"Namsel: An Optical Character Recognition System for Tibetan Text," *Himalayan Linguistics* 15 (1), 2016, pp. 12-30. doi:10.5070/H915129937.

Sawerthal, Anna

"A Newspaper for Tibet: Babu Tharchin and the "Tibet Mirror" (Yul phyogs so so'i gsar 'gyur me long, 1925-1963) from Kalimpong," Heidelberg University Library, 2018. doi:10.11588/heidok.00025156.

Schubert, Johannes

"Typographia Tibetana. Eine Studie über die ausserhalb Tibets verwendeten Typen zum Druck tibetischer Texte." *Gutenberg-Jahrbuch* 25, 1950, pp. 280–98.

Publikationen des modernen chinesisch-tibetischen Schrifttums. Veröffentlichung / Deutsche Akademie der Wissenschaften, Institut für Orientforschung 39. Berlin: Akademie-Verlag, 1958.

Stokes, Peter A., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem
"The eScriptorium VRE for Manuscript Cultures," *Classics@ Journal, Ancient Manuscripts and Virtual Research Environments* (18), 2021. Available online https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (accessed January 14, 2025).

TibSchol
"The Dawn of Tibetan Buddhist Scholasticism (11th-13th c.)," 2022. Available online https://cordis.europa.eu/project/id/101001002 (accessed January 29, 2024).

Wang-Toutain, Françoise
"Base de données et moteur de recherche sur le *Mirror*. Le site Salamandre du Collège de France." *In* Françoise Wang-Toutain and Marie Preziosi (eds.) *Cahiers du Mirror* Paris: Collège de France, 2018, pp. 217–221.

Wylie, Turrell
"A Standard System of Tibetan Transcription," *Harvard Journal of Asiatic Studies* (22), 1959, pp. 261–67.

Appendix: Manual for transcribing historical Tibetan newspapers (in Transkribus)

Transcription should generally follow the generic rule of **What You See is What You Transcribe** (**WYSIWYT**). Normalisation and harmonisation of the corpus will be achieved later, i.e. after OCR/HTR and before NLP.

We are interested in the original text of the newspapers. Therefore, later annotations, handwritten notes, library and other stamps etc., must not be transcribed. The text must be transcribed without correcting spelling or other mistakes.Where the existing Unicode does not provide letters, a compromise was found and followed by the team.

1 Abbreviations

1.1 Kung yig

(1) Abbreviations such as the "reversed T" ཊ་ for abbreviating the final consonants -gs -གས་ should be maintained in the transcription.

(2) When transcribing ume text, transcription in uchen often is difficult or impossible. In general, all abbreviations should be maintained in the transcription. It is best to refer to the available dictionaries to identify abbreviations and reference them in notes to the transcription.

1.2 Abbreviations in languages other than Tibetan (Chinese, English, Hindi)

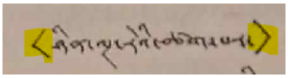
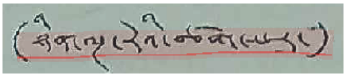
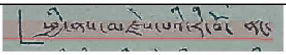
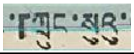
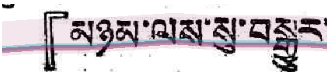
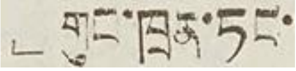
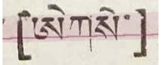
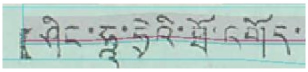
Abbreviations in other languages must be transcribed as in the original.

2 *Transcribing parenthesis, bullet, and punctuation marks*

2.1 *Parenthesises and brackets*

In historical newspapers, a wide range of signs are used for or in the way of quotation marks, parenthesis or brackets:

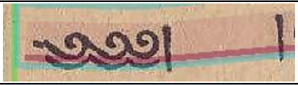
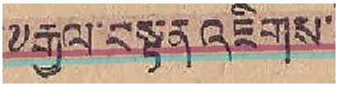
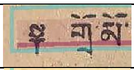
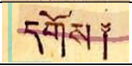
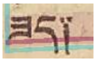
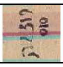
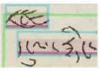
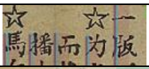
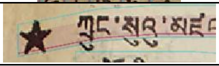
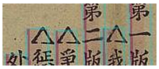
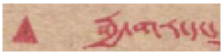
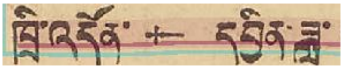
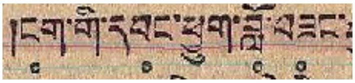
Table 1 Parenthesis and brackets

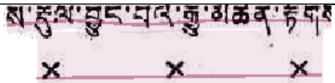
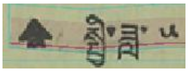
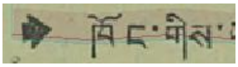
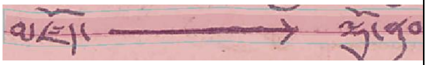
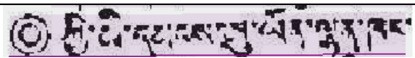
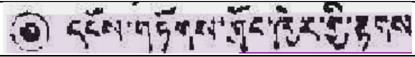
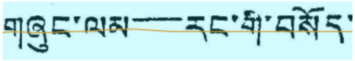
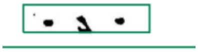
Description	Example	Transcription
angle brackets		< … > U+003C / U+003E
round brackets		(…) U+0028 / U+0029
Chinese traditional single quotation mark	 	U+300C / U+300D
Chinese traditional double quotation mark		『…』 *U+300E / U+300F*
Chinese traditional vertical single quotation mark		﹁…﹂ U+FE41 / U+FE42
square brackets		[…] U+005B / U+005D
lenticular brackets		【…】 U+3010 / U+3011

2.2 *Bullets and punctuation marks*

Tibetan traditionally uses a wide range of bullets and punctuation marks, but also signs that highlight names of highly revered persons (*che mgo*) or point out particular syllables for emphasis or to indicate a second layer of meaning (*ngas bzungs sgor rtags*). In newspapers, additional signs such as stars or triangles are used to indicate enumerations.

Table 2 Bullets and punctuation marks

Name	Example	Unicode
yig mgo mdun ma *yig mgo sgab ma*		༄ U+0F04 ༅ U+0F05
che mgo preceding the names of high incarnates		ྸ U+0F38
sbrul shad		༈ U+0F08
nyis tsheg shad		༐ U+0F10
rin chen spungs shad		༑ U+0F11
gter tsheg		༔ U+0F14
sgra gcan 'char rtags		༗ U+0F17
White Star		☆ U+2606
Black Star		★ U+2605
White Up-Pointing Triangle		△ U+25B3
Black Up-Pointing Triangle		▲ U+25B2
Dagger		† U+2020
ngas bzungs sgor rtags (Emphasis mark)		༷ U+0F37

Name	Example	Unicode
ku ru kha (Iteration mark)		྾ U+0FBE
Upwards Squared Arrow		🠹 U+1F839
Rightwards Squared Arrow		🠺 U+1F83A
Long Rightwards Arrow		⟶ U+27F6
Bullseye		◎ U+25CE
Fisheye		◉ U+25C9
Long dash similar to Em-dash		— U+2014
Dot highlighting page numbers		· U+00B7

3 *Spaces, gaps, and dotted lines*

3.1 *Dotted lines*

In Tibetan typography, the space between the last letter on a line and the end of the line is filled with a dotted line. This dotted line indicates that the statement is not yet finished and continues the following line. Such dots, therefore, have a function different from the inter-syllable *tsheg* (0F0B and the non-breaking 0F0C). In *uchen,* a rounded or triangular dot usually represents both signs. The difference becomes immediately apparent in the above *ume* example. The inter-syllable tsheg is represented in *ume* by a comma-like stroke ་ (similar to the *ume*: ང *nga*, yet slightly shorter).

Problem: No Unicode sign is available for line-filling dots despite their frequent appearance in manuscripts, woodblock prints, and printed texts up to the establishment of computer typesetting for Tibetan (perhaps in the late 1990s?).

Solution: In *ume*, both *Tsheg* and dotted lines are transcribed with a *Tsheg* (which mirrors the use of *uchen*).

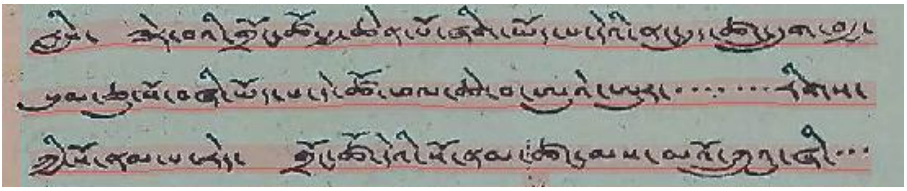


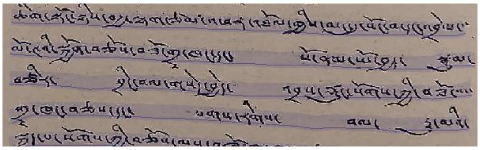
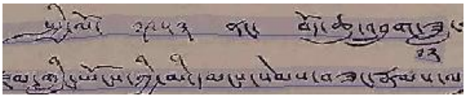
Figure 12 Intersyllabic tsheg and line/space-filling dots in ume

3.2 Gaps and spaces

Tibetan often has gaps of irregular size between statements and between the individual parts of lists. Hence, this gap does not necessarily indicate the end of a statement or sentence.

Besides such gaps, Tibetan usually does not feature "white spaces". However, in newspapers – in comparison to the gap – relatively short and regular "space" can be found, often before and after years.

Table 3 Gaps and spaces

"irregular" gap		SPACE
Short regular space		SPACE

Since Transkribus is unable to differentiate between spaces and tabs, both longer "irregular" gaps and shorter gaps as well as spaces will be transcribed as SPACE (BAR)

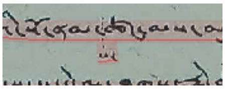
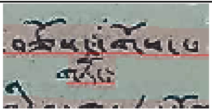
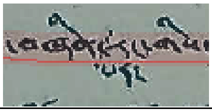
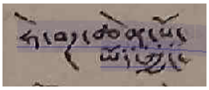
4 *Corrections, additions, marginalia*

4.1 *Corrections of Tibetan text in the form of interlinear additions*

Corrections and additions placed under the baseline are widespread, particularly in handwritten Tibetan texts. For example, where a single letter is missing in a Tibetan syllable, the scribe adds the letter underneath the syllable and draws a fine dotted line indicating exactly where the letter is supposed to be placed by the reader. Such single letters usually are not accompanied by a *Tsheg*.

The insertion of whole syllables, words, or phrases is usually done in the same way, only that a *Tsheg* generally delimits the syllables. However, the dotted line occasionally may be missing, probably when the insertion is supposed to precede the first syllable of a line or a separate statement.

Table 4 Transcribing interlinear corrections and additions

Addition of a single letter		Create a separate baseline for insertion; break the baseline at the dotted line and adjust the reading order accordingly.
Addition of a word	 	Create a separate baseline for insertion; break the baseline at the dotted line, and adjust the reading order accordingly.
Addition of a phrase or sentence		Create a separate baseline for insertion, break the baseline at the dotted line, and adjust the reading order accordingly

4.2 *Marginalia*

Marginalia will be transcribed as a normal paragraph but assigned a marginalia structural tag (see section 5 below), which allows the text region to be identified outside of the letter block.

5 *Tags*

To keep everything simple, we do not want a too large set of tags, which would make the transcription far too complex. The tag set should be limited to the most basic set of tags.

Structural tags are used to mark text regions and classify the information contained. In the Divergent Discourses project, text regions are important not only for information extraction but also for segmenting larger texts into smaller units. Moreover, marking up text regions that contain horizontal Chinese, vertical Chinese, or English text (vs. "unmarked" Tibetan text) will allow extracting text by the respective language.

On the transcription level, several **textual tags** are used that help understand issues on a textual level. For example, they indicate different scripts or languages in the original and mark defects such as tears or stains that influence the legibility of the text. In Transkribus, abbreviations can be tagged, and a model trained to resolve them to their standard form automatically (Griffiths 2022a). In the Divergent Discourses project, however, we decided to perform the normalisation in a separate postprocessing step (Kyogoku *et al.* 2024). While Transkribus provide some tags, others have to be created by each user anew.

5.1 *Structural tags*

Structural tags are the key to information extraction. The tags allow to differentiate and access different regions and their content. It would thus be possible to extract authors' names by extracting only text from the text regions classified CREDIT. In the case of Divergent Discourses, it is important to allow access to the information provided in various languages and scripts, e.g. transcribe the text contained with different models. Consequently, three corresponding tagsets for the languages Tibetan (without prefix), horizontal Chinese (prefixed CNH_), vertical Chinese (CNV_), and English or script in Roman script (ENG_) were created.

Table 5 Structural tag set for annotation of Tibetan newspapers

Tag	Explanation
NEWSPAPER-TITLE	[Tibetan] main title of the newspaper
CNH_NEWSPAPER-TITLE	[horizontal Chinese]
ENG_NEWSPAPER-TITLE	[Latin script/English]
HEADER	[Tibetan] is the top line of a page that usually has the page number, date, and/or name of the newspaper
CNH_HEADER	[horizontal Chinese]
ENG_HEADER	[Latin script/English]
OTHER	[Tibetan] other bibliographic information including the name of the newspaper in other languages, issue number, registration number, dates etc. that is found often underneath the newspaper title or in a separate block on the title page of the newspaper
CNH_OTHER	[horizontal Chinese]
CNV_OTHER	[vertical Chinese]
ENG_OTHER	[Latin script/English]
PARAGRAPH	[Tibetan] main body of the newspaper text
CNH_PARAGRAPH	[horizontal Chinese]

Tag	Explanation
CNV_PARAGRAPH	[vertical Chinese]
ENG_PARAGRAPH	[Latin script/English]
CAPTION	[Tibetan] explanatory text under or next to an illustration (graphical image, photograph, map, etc).
CNH_CAPTION	[horizontal Chinese]
CNV_CAPTION	[vertical Chinese]
ENG_CAPTION	[Latin script/English]
SECTION-HEADING	[Tibetan] heading of a section within a newspaper that is consistently marked with the same heading and features one or more news items, e.g. *khams yul ni*
CNH_SECTION-HEADING	[horizontal Chinese]
CNV_SECTION-HEADING	[vertical Chinese]
ENG_SECTION-HEADING	[Latin script/English]
PAGE-NUMBER	[Tibetan] marks the page number ((under discussion → do we actually need this as the page number information will be recorded in the metadata anyways?))
HEADING	[Tibetan] marks all headings including sub-headings in the newspapers
CNH_HEADING	[horizontal Chinese]
CNV_HEADING	[vertical Chinese]
ENG_HEADING	[Latin script/English]

Tag	Explanation
CREDIT	[Tibetan] marks information to the authorship of a text, image or illustration. This includes also agencies and institutions
CNH_CREDIT	[horizontal Chinese]
CNV_CREDIT	[vertical Chinese]
ENG_CREDIT	[Latin script/English]
MARGINALIA	[Tibetan] marks text printed on the margins or often in the fold of the newspaper
CNV_MARGINALIA	[horizontal Chinese]
ENG_MARGINALIA	[Latin script/English]
CONTINUED	[all languages/scripts] marks indications that the text is continued from/on a preceding/following page or issue
PAGE_NUMBER	[all languages/scripts] marks a page number usually only if the page number is not part of the HEADER
OTHER	[Tibetan] marks text not related to newspaper content, such as publishing information, subscription prices, or announcements by the editors
CNH_OTHER	[horizontal Chinese]
CNV_OTHER	[vertical Chinese]
ENG_OTHER	[Latin script/English]

5.2 *Textual Tags*

In addition to structural tags, Divergent Discourses uses a limited set of textual tags. Most importantly, the UNCLEAR tag allows the marking of illegible parts of the text, which can then be excluded from training. The remaining tags are used to mark different scripts or script styles in the text.

Table 6 Textual tag set for Tibetan newspapers

TAG	**Explanation**
CHIN	Chinese text (left-to-right)
CHIN-VERT	vertical Chinese text
DBU-MED	Tibetan text in cursive script (incl. *dbu med*, *'khyugs* etc)
ENG	text in English language/ Roman script
UNCLEAR	passages that are barely legible due to fading, tears, wholes, stains etc.
BLACKENING	passages that are intentionally blackened

❖