

# Revue d'Etudes Tibétaines

**From Print to Pixels:  
Building Digital Tools for Modern Tibetan Textual  
Analysis**

Edited by  
Franz Xaver Erhard, Robert Barnett,  
and Nathan W. Hill



numéro soixante quatorze — February 2025

# Revue d'Etudes Tibétaines

numéro soixante quatorze — February 2025

ISSN 1768-2959

**Directeur** : Jean-Luc Achard.

**Comité de rédaction** : Alice Travers, Charles Ramble, Marianne Ginalski, Jean-Luc Achard.

**Comité de lecture** : Ester Bianchi (Università degli Studi di Perugia), Fabienne Jagou (EFEO), Per Kværne (University of Oslo), Rob Mayer (Oriental Institute, University of Oxford), Fernand Meyer (CNRS-EPHE), Françoise Pommaret (CNRS), Ramon Prats (Universitat Pompeu Fabra, Barcelona), Charles Ramble (EPHE, CNRS), Françoise Robin (INALCO), Alice Travers (CNRS), Jean-Luc Achard (CNRS).

**Périodicité**: La périodicité de la *Revue d'Etudes Tibétaines* est généralement bi-annuelle, les mois de parution étant, sauf indication contraire, Octobre et Avril. Les contributions doivent parvenir au moins six (6) mois à l'avance. Les dates de proposition d'articles au comité de lecture sont Novembre pour une parution en Avril, et Mai pour une parution en Octobre.

**Participation**: La participation est ouverte aux membres statutaires des équipes CNRS, à leurs membres associés, aux doctorants et aux chercheurs non-affiliés. Les articles et autres contributions sont proposés aux membres du comité de lecture et sont soumis à l'approbation des membres du comité de rédaction. Les articles et autres contributions doivent être inédits ou leur réédition doit être justifiée et soumise à l'approbation des membres du comité de lecture. Les documents doivent parvenir sous la forme de fichiers Word, envoyés à l'adresse du directeur ([jeanluc.achard@sfr.fr](mailto:jeanluc.achard@sfr.fr)).

**Comptes-rendus**: Contacter le directeur de publication, à l'adresse électronique suivante : [jeanluc.achard@sfr.fr](mailto:jeanluc.achard@sfr.fr)

**Langues** Les langues acceptées dans la revue sont le français, l'anglais, l'allemand, l'italien, l'espagnol, le tibétain et le chinois.

La *Revue d'Etudes Tibétaines* est publiée par l'UMR 8155 du CNRS (CRCAO), Paris, dirigée par Matthias Hayek.

**Hébergement**: <http://www.digitalhimalaya.com/collections/journals/ret/>





# Revue d'Études Tibétaines

*numéro soixante quatorze — February 2025*

## **From Print to Pixels: Building Digital Tools for Modern Tibetan Textual Analysis**

Edited by

Franz Xaver Erhard, Robert Barnett,  
and Nathan W. Hill

### **Robert Barnett and James Engels**

Assembling a Digital Toolkit: An Introduction to Text-  
mining for Modern Tibetan

pp. 5–43

### **Franz Xaver Erhard**

The Divergent Discourses Corpus: A Digital Collection  
of Early Tibetan Newspapers of the 1950s and 1960s

pp. 44–80

### **Christina Sabbagh**

Enhanced ATR Accuracy for Tibetan Historical Texts:  
Optimising Image Pre-processing for Improved  
Transcription Quality

pp. 81–127

### **Franz Xaver Erhard**

Text and Layout Recognition for Tibetan Newspapers  
with Transkribus


pp. 128–171

- Franz Xaver Erhard und Xiaoying 笑影**  
Foreign Names and Places in Tibetan Newspapers  
of the 1950s and 1960s pp. 172–186
- Yuki Kyogoku, Franz Xaver Erhard, James Engels, and  
Robert Barnett**  
Leveraging Large Language Models in Low-resourced  
Language NLP: A spaCy Implementation for Modern  
Tibetan pp. 187–220
- Ronald Schwartz and Robert Barnett**  
Religious Policy in the TAR, 2014–24: Topic Modelling  
a Tibetan-Language Corpus with BERTopic pp. 221–260
- James Engels and Robert Barnett**  
Developing a Semantic Search Engine for Modern  
Tibetan pp. 261–283
- Natalia Mikhailova**  
*The Tibet Mirror*, “Friends” of Tibet, and the Inter-  
nationalisation of the Tibet Question pp. 284–328



# Assembling a Digital Toolkit: An Introduction to Text-mining for Modern Tibetan

Robert Barnett (SOAS University of London)  
and  
James Engels (University of Edinburgh)

he rapid development of new computational tools and methodologies for use in the humanities and social sciences raises both hopes and challenges for those who try to work with them. In this introduction to this special issue “From Print to Pixels: Building Digital Tools for Modern Tibetan Textual Analysis” of the *RET*, we describe some of the emerging tools designed to assist researchers working in the digital humanities on texts in the Tibetan language and discuss some of the challenges that are part of that effort. Our discussion complements the paper by Meelen, Nehrdich and Keutzer (2024), which described field-wide developments in a wide range of computational tasks involving Tibetan, including Handwritten Text Recognition (HTR), post-processing correction, speech-to-text technology, and digitally-assisted language learning for Tibetan. By contrast, this special issue presents, through a set of seven studies, a step-by-step summary of the main tasks involved in developing a single digital humanities project in the Tibetan studies field.

Our project, “Divergent Discourses,”<sup>1</sup> aims to identify the various discourses and narratives produced in newspapers within Tibet (after

---

<sup>1</sup> The Divergent Discourses project received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (<https://gepris.dfg.de/gepris/projekt/508232945?language=en>), and from the Arts and Humanities Research Council (AHRC) under project reference AH/X001504/1

1950, in China) and Tibetan print media produced outside Tibet (mainly in India or Nepal) in the late 1950s and early 1960s. The project aims to track the historical evolution of discourses on both sides of the Himalayas during such pivotal episodes as the “democratic reforms”, the various uprisings by Tibetans in the later 1950s, the flight to India, and the ensuing “elimination of the rebellion” within Tibet. To do this, the project team has worked for the last two years on developing or refining a number of computational tools and processes that are designed to facilitate text-mining, and which we have adapted for use with modern Tibetan texts. Developing those tools has involved a number of tasks, which we briefly describe in this introduction. These involved primarily (1) finding, collecting and scanning the source documents; (2) checking and improving the images; (3) training a machine to transcribe the image content into textual form; (4) post-processing, including normalisation, paragraph extraction and adding metadata; (5) training a Tibetan-language model to enable the text-mining platform to work with Tibetan texts; (6) developing or adapting tools to recognise words in Tibetan and their parts-of-speech (POS) in order to produce training data for the language model; and (7) developing a set of text-mining tools that rely on a technology that is distinct from that used by the text-mining platform, in case the latter failed to work well with Tibetan.

Each of the papers in this special issue of *RET* describes some of the choices, considerations and (eventually) successes that we faced in carrying out these tasks. In addition, this introduction includes brief sketches of the historical background of these processes and the rapid changes taking place in the field of computationally-assisted text mining, together with the implications of these changes for digital Tibetan studies. The papers also include a description of the Divergent Discourses corpus of early Tibetan newspapers and their history, a study of Tibetan transcribing practices for foreign names, an analysis of religious policy in Tibet in the last decade, and a survey of a

---

(<https://gtr.ukri.org/projects?ref=AH%2FX001504%2F1>). For more information on Divergent Discourses, see <https://research.uni-leipzig.de/diverge/> (accessed on January 10, 2025) and the other contributions to this special issue.

particular discourse in an overseas Tibetan newspaper in the 1950s and early 1960s. We hope these papers will be helpful as a kind of travel guide to other researchers setting out on similar journeys in the future.

### 1 *Building the Corpus*

The project's source materials are 16 Tibetan-language newspapers published in the 1950s and the early 1960s. These add up to 16,718 newspaper pages. Although still a comparatively small collection, it would be very hard for a single researcher or a small team to study this volume of texts unless assisted by digital tools that allow automated or semi-automated textual analysis. It is not enough, however, just to acquire such tools from the internet and then apply them to the texts in our collection. Before any such application or analysis can be performed, a number of preparatory tasks are necessary, including, in particular, training these tools and their underlying platforms or systems so that they are able to work with texts that are in the Tibetan language, and specifically, in our case, in modern Tibetan. Only then can we reliably use computational tools to identify keywords, topics, sentiments, names and other details in the texts, which in turn will enable us to trace in detail the narratives and discourses circulating at the time when these articles were published.

Developing and adapting these tools involves numerous decisions and value judgements, most of them requiring considerable technical expertise.<sup>2</sup> In this introduction, we summarise some of the factors involved in those decisions and the basic concepts underlying them. As many researchers in the field of Tibetan studies know well, these concepts are largely drawn from the field known as Natural Language Processing (NLP). They include, firstly, the notion of a corpus, which means a large number of machine-readable texts compiled into a single body, enabling what is popularly called "big data" analysis and text-

---

<sup>2</sup> The project's technical decisions were guided and implemented primarily by the team's technologists and research assistants, James Engels and Christina Sabbagh in London and Yuki Kyogoku in Leipzig.

mining. The composition of the Divergent Discourses Corpus is described in the paper in this issue by **Franz Xaver Erhard**, “The Divergent Discourses Corpus: A Digital Collection of Early Tibetan Newspapers of the 1950s and 1960s” (Erhard 2025a). Erhard provides extensive historical background about the emergence of Tibetan-language newspapers since the early 1900s and describes the rapid increase in the publication of such papers on both sides of the Himalayas during the 1950s. Noting their exceptional importance in the case of Tibetan studies, not least because access to archives in Tibetan areas of China is rarely permitted for foreign (or even local) researchers, he also shows how our collection, like any corpus, contains structural biases as a result of incomplete collections, lost items, illegible copies, and so forth. Nevertheless, the corpus represents the first known attempt to make a comprehensive set of Tibetan-language newspapers from that era publicly available.

Creating a corpus is more or less an essential requirement for applying the various tools or techniques of text-mining that are so far available. Obviously, the main task in corpus creation is acquiring copies of each page of each newspaper, which might initially take the form of photographs, microfilm, or microfiche, and then, if necessary, converting any chemically-produced (sometimes incorrectly termed analogue) images into a digital format by scanning them to produce a digital image file. But a number of operations – usually referred to as “pre-processing” – have to be carried out on each image before it can be added to a corpus. Among these is improving the quality of any damaged or sub-standard images by using image-enhancement tools. As Erhard noted in his paper, sub-standard images can bias the representativeness of a corpus, leading researchers to assume that their findings reflect the totality of the corpus contents when, in fact, they may only include those with sufficient image quality to be accurately analysed and transcribed by a machine.

It was to reduce this risk that **Christina Sabbagh** developed a procedure for assessing and improving the quality of images in the corpus. As she explains in her paper, “Enhanced HTR Accuracy for Tibetan Historical Texts – Optimising Image Pre-processing for Improved Transcription Quality” (Sabbagh 2025), this task was



particularly necessary in the case of a microfilm edition published by the China National Microforms Import and Export Corporation, Beijing, in the 2000s. That microfilm edition contains the only known copies of several thousand newspaper pages in Tibetan, but many of its images are missing, incomplete, underexposed, stained, or obscured by dark patches. She describes her methodology for automating the identification of those damaged images, and, in some cases, for digitally improving them.<sup>3</sup> She notes that no single procedure can fully restore all poor-quality images due to their varying levels of degradation, but enough to recover and extract the majority of the text contained. Sabbagh suggests additional post-processing strategies could be explored in the future to try to compensate for these difficulties, including the Turing Institute's MapReader tool and the Impreso Project's keyword suggestion tool (see also the description of the Norbu Ketaka project in Luo and van der Kuijp 2024, which uses NLP and computer vision models to “clean up” or improve machine-transcribed Tibetan texts). Sabbagh emphasises the importance of recognising that automatic text recognition is prone to errors, and that downstream applications using machine-readable text must be designed with this assumption in mind to yield truly representative conclusions.

The next major step towards a machine-readable corpus is automatic transcription or “text recognition” – using a machine to extract visual information, such as letters and words, from an image file and to convert that information into textual form. The textual form, in this case, means a file that can be read by a machine such as a word processor, so that it can be processed and manipulated by computational tools. This process is conventionally known as “optical character recognition” or OCR. In his paper, “Text and Layout Recognition for Tibetan Newspapers with Transkribus” (Erhard 2025b), **Franz Xaver Erhard** points out that strictly speaking automatic transcription cannot be termed “OCR” when it involves Tibetan texts, because in the case of Tibetan the transcribing algorithms are trained

---

<sup>3</sup>. The project's image enhancement code is at [https://github.com/DivergentDiscourses/dd\\_custom\\_preprocess](https://github.com/DivergentDiscourses/dd_custom_preprocess) (see also Sabbagh *et al.* 2024a, b).

to recognise the shapes created by strings of letters (usually a line) on the page, not the individual characters. This shape-recognising technology, known as Handwritten Text Recognition or HTR, was first discussed in more detail in the Tibetan context in Griffiths 2024.<sup>4</sup> In his paper, Erhard describes the complex and time-consuming task of training an HTR machine to transcribe Tibetan newspaper texts. One first has to select an automatic transcription programme, in this case, the online service Transkribus,<sup>5</sup> and then produce what is called “Ground Truth”, meaning hundreds of pages of accurate and verified manual transcriptions that can be fed to the machine alongside the digital images of those pages. These train the HTR algorithm to recognise the specific handwriting or typeface of one or other publication or source. For best results in model training, the Ground Truth used for training must adhere to a standard set of transcribing principles to ensure that all pages are transcribed identically. While this transcription imperative is easily accepted in theory, in practice many challenges arise.

One of these challenges involves Tibetan punctuation practices. The Tibetan writing system uses a number of unique symbols to mark texts, such as the “fish-eye”, the “bullseye”, the “black up-pointing triangle”, or the *che mgo* preceding the name of high incarnate lamas. In addition, Tibetan newspapers sometimes incorporate symbols borrowed from other writing systems such as Chinese or English. The transcription model, therefore, has to be instructed as to which Unicode character should be used to transcribe each of these symbols or marks. Details of the transcribing conventions developed by the project are given by Erhard in his “Manual for transcribing historical Tibetan newspapers (in Transkribus)”, included as an appendix to his paper on “Text and Layout Recognition”.

In the case of modern printed books, HTR is a relatively straightforward task because the physical layout of a page in such a

---

<sup>4</sup> Erhard uses the term Automatic Text Recognition (ATR) as an umbrella term including both OCR and HTR.

<sup>5</sup> For more on Transkribus ([www.transkribus.org](http://www.transkribus.org)), see Kahle *et al.* 2017.

book is usually standardised.<sup>6</sup> But, as Erhard shows in his article, this is not the case with early Tibetan newspapers, which have highly complex layouts with varying numbers of columns as well as photographs, advertisements, mastheads, headings, sub-headings, captions, page numbers, and other forms of design or layout features on a page (these are known as “text regions”). A transcription model has to be trained to recognise these regions and to follow the different reading conventions needed to correctly transcribe texts within each of them. To teach these conventions to the machine, Erhard had to develop what is known as a “Field Model” in Transkribus, an algorithm trained to recognise the different text regions, as well as to train models that would recognise what are called “baselines”, the line which, conceptually, runs through the centre of each line of text, and “line polygons”, which are boxes that one teaches the machine to recognise as tightly enclosing each line of text. In his paper, Erhard describes the innovative three-step workflow he developed to enable automatic text recognition to compensate for the complex layouts of Tibetan newspapers. It took some two years of work, but with this method, he was able to successfully train a Transkribus model to read early Tibetan newspapers for the Divergent Discourses project. Whereas Transkribus advises that transcription models with a Character Error Rate of 10% or less should be considered successful, our model, TibNews4All 0.2, has a Character Error Rate of 2.52%.<sup>7</sup>

## 2 After the HTR stage: Post-processing

Once the text recognition process has produced digital transcriptions of the collected images with an acceptable level of accuracy, a number

---

<sup>6</sup> The Divergent Discourses project published a model on Transkribus for modern printed Tibetan books from the PRC in March 2024, Tibetan Modern U-chen Print 0.1 (TMUP 0.1). It is the first Transkribus HTR model for printed Tibetan language publications in Uchen (འུ་ཅེན་ *dbu can*) script, available at <https://readcoop.eu/model/tibetan-modern-u-chen-print/>; see also Erhard *et al.* 2023.

<sup>7</sup> The TibNewsOne4All 0.2 (ID 169581) is available in Transkribus at <https://www.transkribus.org/model/tibnewsone4all> (accessed on January 14, 2025).

of post-processing steps have to be taken to maximise the value of the corpus for researchers. The first of these is normalisation. This step, which is carried out after the transcription process on a copy of the raw transcriptions, involves writing code that standardises variations in orthographic or other practices in the document. In our case, we wrote code that included ensuring that all signs and characters in the text are encoded according to the Unicode UTF-8 standard, spelling out abbreviations, replacing any non-Unicode characters, standardising numbers, and removing non-significant spacing or *tshegs* (the small inline dot or inverted triangle (·) used to separate syllables).<sup>8</sup> Without this process, searches and similar tools will fail to capture all, or at least most, instances of any given term or feature – searching for the number “8”, for example, would find only texts which used the Arabic numeral 8, and not those that use the Tibetan numeral ༘. A well-designed corpus will usually retain an option for researchers to view the original or “raw” form of each transcribed text in case they wish to see the spellings, numbers or punctuation used in the transcribed text prior to normalisation.<sup>9</sup>

In our case, the normalisation process brought our attention to a specific issue in Tibetan orthography: the wide range of variation in the spelling of non-Tibetan words and names. This led to the paper by **Franz Xaver Erhard** and **Xiaoying**, “Foreign Names and Places in Tibetan Newspapers” (Erhard & Xiaoying 2025), which presents an innovative study of Tibetan naming practices in the 1950s and 1960s. It shows the remarkable variation in spellings of foreign terms in Tibetan texts at that time, a reflection, the authors note, of the rushed nature of China’s translation project in Tibet, where seemingly officials had no time to set up bodies in their new territory to standardise such translation practices. Erhard and Xiaoying show the wide variations in spellings that this produced and explain some of the reasons for those discrepancies – they found, for example, twelve forms in Tibetan of the name of the former Chinese premier, Zhou Enlai, many of them related

---

<sup>8</sup> For our code for post-transcription normalisation, see Kyogoku *et al.* 2024.

<sup>9</sup> The project’s normalisation process is discussed in Kyogoku *et al.* 2025, section 3.3. For the code, see <https://github.com/Divergent-Discourses/TibNorm>.

to differences in pronunciation practices in different Tibetan or Chinese dialects. Again, these variants would ideally be normalised if a researcher is going to be able to find all instances of a given name in the corpus.<sup>10</sup>

In some cases, the texts to be included in a corpus do not need automatic transcription because they are “born digital” – that is, they already exist in a machine-readable form, such as those that are found on the internet in HTML format. This is rare in the case of historical documents, but even with texts that can be directly downloaded or

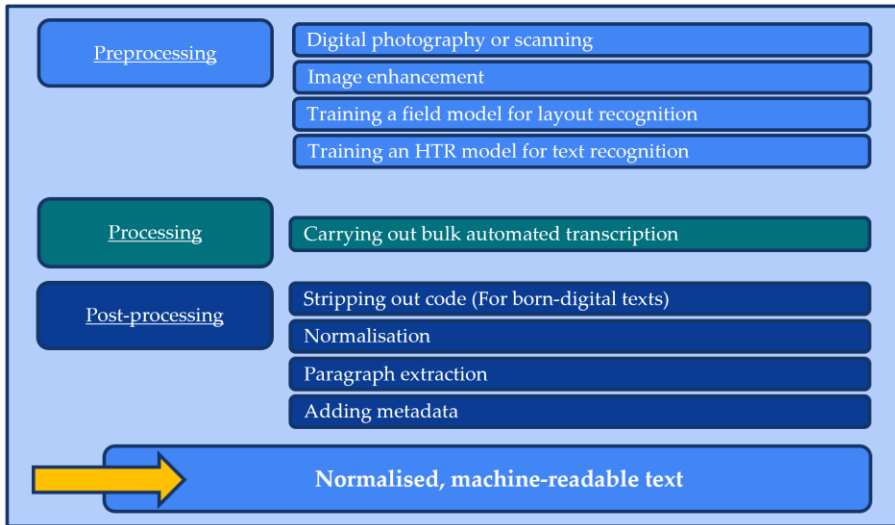


Figure 1 The basic steps required to produce normalised, machine-readable text.

“scraped” directly from websites in such a format, one then needs to invoke a code or procedure that will strip out non-essential features like formatting marks, non-standard characters, or HTML coding. Such code is, however, already available in many forms online.<sup>11</sup>

An essential step in the pre-processing pipeline, before a corpus can be used effectively, is the addition of metadata to each item in the collection, including the title of the source publication, the date and

<sup>10</sup> The name-lists compiled by Erhard and Xiaoying are available at <https://zenodo.org/records/14526125>. See Erhard, Xiaoying *et al.* 2024.

<sup>11</sup> See for example Leonard Richardson, “beautifulsoup4 4.12.3”, <https://pypi.org/project/beautifulsoup4/> (accessed January 21, 2025).

place of publication, and the number of the page on which it was originally published. Basic metadata is attached by Transkribus to each transcribed text in an XML file that it outputs together with the transcription. This includes the file name of the source and the tags for identified text regions. For our project, we initially wrote code that aimed to add metadata extracted from the online catalogues of the libraries which had held the original copies of the newspapers, but for technical reasons this failed and we instead added such metadata by hand. In addition, because Tibetan texts do not consistently mark sentences, we decided to treat the paragraph as our smallest unit of analysis and so wrote code that subdivided each transcribed text into paragraphs or segments of text. This code attached metadata to each of these paragraphs, linking each to its source document.<sup>12</sup>

### 3 *Preparing Tools for Text-analysis*

After the pre-processing stages, the texts can be grouped into a collection or corpus, which can be simply a folder on a hard drive containing the raw texts. The process of developing or applying tools for analysis of that corpus can then begin. There are four main analytical tools or methods which are commonly used in corpus-based text-mining:

- **keyword searches**, where the computer searches for occurrences of a particular word or phrase (a “string” in computer-speak) and computes their frequency;
- **sentiment analysis**, where the programme identifies texts that share a particular sentiment, such as happiness or disapproval;
- **topic modelling**, where the computer recognises which passages focus on a particular topic and groups paragraphs or sentences into shared topics or themes; and
- **named entity recognition (NER)**, which identifies the elements in a text that are names of entities such as a person, institution or place.

---

<sup>12</sup> See [https://github.com/Divergent-Discourses/transkribus\\_utils](https://github.com/Divergent-Discourses/transkribus_utils) and Engels *et al.* 2024.

Over the last four decades or so, computational linguists have developed four main approaches or techniques for carrying out these tasks. These approaches have evolved in terms of sophistication and complexity. The first approach, which we can term traditional, is rule-based: the computer is taught a set of rules that roughly replicates the grammatical features of a given language. Initially, the rules developed for the purposes of this approach were of a static or first-order type of complexity, meaning that they accounted only for basic elements of a language, such as “a space marks a word” or “a final -s marks a plural”. In time, rules of a second-order or dynamic type were developed to address more complex conditionalities, such as recognising “Great” and “Britain” as one word. These rules provide the computer with a basic model for the language of the texts in its corpus.

From the early 1990s, developers began to introduce a second approach, based not on rules but on machine learning. This approach, also known as supervised learning, which is now also considered traditional, involved feeding large quantities of hand-annotated (“marked up”) texts to the computer. The computer would then transform the text into numerical representations based on hand-selected features (like word counts or grammatical tags) and learn to recognise patterns in the language through probability assessment used by those texts.

The third approach, which emerged in the early 2000s, also involves machine learning, but is based on the use of vectors. Vectorisation (often called “embedding”) is the term used by computational linguists for the process where texts or words are transformed into numbers or numerical strings and then analysed mathematically. In most cases, this approach again requires large amounts of pre-annotated training data for each given language.<sup>13</sup> As with the previous two approaches, this method requires the computer to be

---

<sup>13</sup> Approaches of this type that are known as “unsupervised learning” allow the computer to identify patterns without pre-labelled examples, so they do not require pre-annotated training data.

provided with a model of each language used by the texts it works with.

Fourthly, since 2017, a new machine-learning approach has been developed which uses what is called “transformer” architecture or technology.<sup>14</sup> The transformer breakthrough resulted from the publication of an article that introduced a feature known as the “attention mechanism” (Vaswani *et al.* 2017). This innovation has revolutionised the field of vectorisation, and thus of NLP. The transformer process works firstly by reducing all the words in a text or corpus to unique numerical strings, one for each word (or token) and sometimes one for each paragraph or document, and then uses the complex algorithms which make up this mechanism to guess with a high degree of accuracy, if the words have been supplied to a computer in sufficiently large quantities, which word or token is most likely to follow or precede any given one.

This has produced the technologies now termed “generative AI”, which have led, among a number of outcomes, to “large language models” (LLMs). These systems are again based on vectorisation, but have greatly enhanced sensitivity to local context within a text or sentence. Transformers, much as with earlier forms of machine learning, require very large quantities of training data – but they do not require that data to have been pre-annotated, and they do not need to know what language is used by the texts they are analysing or to be given a language model in advance.

These four approaches are often combined – rule-based methods can be integrated with machine learning, for example – to enhance accuracy, efficiency, and other criteria. The rule-based approach can be demonstrated by a basic corpus- or text-analysis tool such as Antconc.<sup>15</sup> Antconc can be run on a local drive, and it can use simple

---

<sup>14</sup> Broadly speaking, machine learning refers to an algorithm or sequence of algorithms that can be trained and can then make predictions or judgments about unseen data. This includes everything from a basic conditional probability calculation to black-box transformers like GPT-4.

<sup>15</sup> Antconc is freely available at <https://www.laurenceanthony.net/software/antconc/>. Using it does not require any technical knowledge. To use it for Tibetan, go to “Global Settings/Token Definition”, select “Use Following Definition”, paste





with non-programmers in mind, with a relatively straightforward user interface, it wraps together a number of useful but otherwise largely inaccessible NLP functions. Specifically, the iLCM's most useful features for us are interfaces for full-text keyword searches and metadata storage for individual documents in corpora. It also analyses word frequency and can carry out co-occurrence detection (especially useful for detecting slogans or jargon), topic modelling, and semantic volatility analysis (changes in word meanings over time). The iLCM includes tools for annotating corpora or coding sections of text and so is especially helpful for those using Qualitative Content Analysis on texts. It can also display results in visual or graphic forms, such as network diagrams or charts showing the distribution of a term or topic over time (sometimes called "dynamic topic modelling"), which facilitates diachronic interpretation of a text or corpus.

#### 4 *Developing a Tibetan Language Model*

Like all such programmes, the iLCM is not a stand-alone application: it runs on a particular platform or software package which renders incoming data intelligible and that underlying platform has to be trained to work with Tibetan. In the case of the iLCM, this underlying platform is a popular NLP software library called spaCy.<sup>18</sup> However, spaCy does not include Tibetan among its default-supported languages. So, before being able to use the iLCM to analyse Tibetan texts, we first had to develop a Tibetan language model for use with spaCy. To develop such a model requires preparing large quantities of pre-annotated training data. These texts have to have been cleaned up (extraneous code and so forth has to be removed) and every word in the texts, and every sentence or phrase (or "utterance") in that text, has to have been tokenised – that is, they should be marked as distinct from a previous or subsequent word or utterance. This is the process known as "segmentation" or, more commonly, as "tokenisation". In

---

<sup>18</sup> See <https://spacy.io/>. SpaCy is a general-purpose software package for end-to-end NLP applications, including word segmentation, topic modelling and NER.

some cases, and in particular for training a language model for a platform such as spaCy, training data also has to be marked up with tags that identify the part of speech of each word (“POS tagging”).<sup>19</sup>

Considerable quantities of such training data already exist for Tibetan, prepared and annotated by previous projects.<sup>20</sup> Most of that data, however, consists of texts in classical Tibetan, and a model trained with these will produce numerous errors or omissions when used to read modern Tibetan texts. For our language model, we therefore needed to produce large quantities of annotated data in modern Tibetan. Without these, spaCy would struggle to recognise what is a word or meaningful particle in Tibetan and our analysis tool, the iLCM, would be more or less unable to identify frequencies, topics, entities and other features within the corpus.

In the interim, while developing our annotated training data for modern Tibetan, we carried out a test to see if a makeshift Tibetan-language model using spaCy would enable the iLCM to work with Tibetan. We called this test model “Tibetan for spaCy 1.1” (see Kyoguku *et al.* 2025 in this issue, section 5). We used the training data produced by Dakpa *et al.* (2021a, b) – it consists of only 13 megabytes

---

<sup>19</sup> Some NLP procedures work better if they have also been trained to recognise the “lemma” or root form of each word and the syntactic structure of the language, though for our project we have so far not found it necessary to apply lemmatisation or to add syntactical information to our texts.

<sup>20</sup> Classical Tibetan corpora include, for example, the Asian Classics corpus (available at <http://resources.christian-steinert.de/download/acip-release6-wylie.zip>); the BDRC Corpus, available at <https://zenodo.org/records/821218#.Xu5IOOdYxld> (Wallman *et al.* 2017), annotated as the “ACTib” corpus (<https://zenodo.org/records/821218#.Xu5IOOdYxld>; see Meelen and Roux 2020); the “Tibetan in Digital Communication” corpus (<https://zenodo.org/records/574878>; see Hill & Garrett 2017); the “Lexicography in Motion” corpus (<https://zenodo.org/records/4727108>; see Faggionato *et al.* 2021); and the OpenPecha corpus (<https://github.com/OpenPecha/openpecha-catalog>). Corpora of modern Tibetan texts include the “Nanhai corpus” produced by Esukhia (<https://github.com/Esukhia/Corpora/tree/master/Nanhai>), probably 2017; the Fudan NLP Tibetan Classification corpus, produced by the Natural Language Processing Laboratory of Fudan University, probably 2017 (<https://github.com/FudanNLP/Tibetan-Classification>); and the “Modern Tibetan Corpus” produced as part of “Lexicography in Motion” (see Dakpa *et al.* 2021a, 2021b).

of annotated modern Tibetan texts, a very small amount by NLP standards – which was already available in CoNLL-U, a format which spaCy's default models can natively interpret and learn from. We then added a space after each *tsheg* in the data, fed the training data into spaCy, and instructed it to treat the input data as if it were English. SpaCy thus used its English language model to read the data, interpreted the space after each Tibetan token as indicating the end of an English word, and so added these (actually Tibetan) “words” (actually syllables) to its inbuilt English lexicon. This method produced numerous errors of tokenisation and so forth, but it worked: it enabled the iLCM to read and process the Tibetan documents in our test corpus, as well as to show topics, relational distributions of words and phrases, and even political slogans (Engels *et al.* 2023).

This interim model thus worked as a temporary measure to test the feasibility of using the iLCM, but, since it used a “phoney” tokenisation method that produced frequent errors, it could not be used for any serious textual analysis. To achieve a durable language model for spaCy and hence for the iLCM – or for any pre-transformer approach to NLP and textual analysis involving modern Tibetan texts – we needed to produce sufficient quantities of annotated training data in modern Tibetan to feed to the underlying platform used by our corpus-analysis programme. Only then would that programme be able to search a corpus for particular Tibetan terms, concepts, topics, names or semantic features of interest to the user.

### 5 *Tokenising: Rule-based vs. Transformer-based approaches*

To produce training data consisting of accurately annotated modern Tibetan texts, we needed to find tools that could efficiently carry out tokenisation and POS tagging on such texts. The result of our search for such tools, which enabled us in time to develop a fully-functioning Tibetan language model for spaCy, is described in the paper in this issue by **Yuki Kyogoku, Franz Xaver Erhard, James Engels** and others, “Leveraging Large Language Models in Low-resourced

Language NLP: A spaCy Implementation for Modern Tibetan” (Kyogoku *et al.* 2025).

Tokenisation can be carried out by a basic NLP tool like Antconc just by differentiating lexical units by spaces, punctuation, or other signs; it can then count the tokens, organise them, compute their frequency, and so forth. But tools or procedures that carry out more advanced forms of NLP analysis, like the iLCM and spaCy, need to be able to recognise features of the language they are processing with more precision. In particular, they need to be able to recognise individual words, including polysyllabic ones, as an initial step before any further analysis can be done. In the early stages of NLP, such tools were trained to recognise separate words and utterances by using a rule-based approach to tokenisation. These early tools were designed for use with European languages, mainly English, and so their rules were over-tuned to the specific grammatical requirements of European languages, typically using white spaces to split raw text, along with a basic punctuation set plus apostrophes.

In Tibetan, however, where each syllable is separated by a *tsheg*, only monosyllabic words are marked as distinct. Polysyllabic words are not marked and thus would be invisible to our corpus-analysis tool if our Tibetan texts were tokenised simply by assuming every syllable before a *tsheg* is a word. This would lead, for example, to a collocation such as བོད་རང་སྐྱོང་ལྗོངས་ (Tibet Autonomous Region) being defined as three words, rather than as a single one, albeit with three *tshegs*, four syllables and three recognisable words in the string. Similar difficulties arise with the complex forms of noun and verb morphologies in Tibetan. Rule-based methods of tokenisation for Tibetan are based on Tibetan syntax or on checking words in dictionaries (a process known as “dictionary look up”), but this tends to bias longer words even if the component units of that word should at times be parsed as shorter words positioned consecutively. Tokenisation for Tibetan hence requires a more sophisticated approach, one which can follow second-order rules that address the context of words or that can do probability calculations using a machine-learning approach.

At least one such tokeniser exists: Botok. It is a product of collaboration between the Buddhist Digital Resource Center (BDRC)

and other organisations such as OpenPecha and Esukhia, a nonprofit that specialises in developing digital resources related to Tibetan languages and their textual traditions. Unlike previous tokenisers produced through this collaboration that were purely rule-based,<sup>21</sup> Botok's tokenisation procedure involves first splitting on the *tsheg*, so that every syllable is isolated in sequence. Combinations of syllables are then evaluated using an internal dictionary search, and decisions about what classes of words to search are made based on statistical evaluations of preceding contexts. In addition, Botok is context-dynamic: it has the capability to edit the rules internally when its input data exhibits a repeating but previously unlearned pattern. Importantly, Botok's internal dictionary can be modified or updated. Furthermore, it has the advantage that it is consistent: once its settings have been adjusted to the user's needs, it will tokenise any set of morphemes in the same way.

A number of other approaches have been developed for tokenising Tibetan. These include a machine-learning approach known as a "Memory-based tagger" used by Meelen & Hill (2017) to produce an archive of high-quality annotated training data, and the combined rule-based, memory-based, and deep-learning method used by Meelen, Roux and Hill (2021) to annotate their "ACTib" corpora (see also Faggionato, Hill and Meelen 2022). However, these projects used classical Tibetan texts for their training data and so their tokenisation methods are not attuned to modern vocabulary, particularly compounds such as བོད་རང་སྐྱོང་ལྗོངས་ (Tibetan Autonomous Region), ལྷན་ཁྲིམས་ལྷན་ཁུངས་ (committee) or གུང་ཁག་ (CCP). We therefore turned to the newly emerging transformer-based technologies to explore their capabilities in tokenisation for Tibetan. Using probability-based calculations, these technologies can infer which syllables should be treated as a word or token. To do this, they require very large quantities of training data in the given language, but that training data does not need to have been pre-annotated: the most recent transformer-based models can infer all

---

<sup>21</sup> Botok is built from a tokeniser called PyBö, developed by a collaboration between Esukhia and the BDRC and completed in 2021, although Botok has a larger training dataset than PyBö. See <https://github.com/OpenPecha/Botok>.

necessary information about the language in their training data based solely on pattern recognition in that data. As a result, if they have been given data of sufficient quality, these models are now developing the ability to carry out essential tasks, including tokenisation, on texts where they have not been given information in advance about the language of those texts.<sup>22</sup>

A prominent language model of this kind, developed by Google in 2018, is called BERT (“Bidirectional Encoder Representations from Transformers”). BERT uses transformer architecture to predict what text might come before and after other text. This form of machine learning, however, has to be trained on very large amounts of raw text, which presents a problem for a relatively low-resource language like Tibetan. Nevertheless, a number of applications of BERT have been developed for use with Tibetan by scholars in Tibet or China, and some of these have been made publicly available, including a BERT model for Tibetan called TiBERT (Sun *et al.* 2022) and one called Tibetan BERT (Zhang *et al.* 2022).<sup>23</sup> However, BERT was designed to help computers analyse the meaning of a word based on the words surrounding it, an approach which is applicable for tasks involving classification, such as

---

<sup>22</sup> However, because transformer models often rely on shared linguistic structures from related languages and then transfer their learning to the unseen language, they perform less well on unseen languages that do not share characteristics with those on which they have previously been trained.

<sup>23</sup> The two models mainly differ in size and computational complexity: like BERT Base, TiBERT contains 12 transformer blocks, 768 hidden dimensions, 12 self-attention heads, and 110 million parameters (available at <http://tibert.cmli-nlp.com/>). Tibetan BERT is a scaled-down version of BERT with a decreased computational load, focusing on minimal decrease in performance for use in situations with heavier resource constraints. Tibetan BERT has 4 transformer blocks, 256 hidden dimensions, 4 attention heads, and does not report its parameter figure. TiBERT is available through the creators’ own distribution ([https://huggingface.co/UTibetNLP/tibetan\\_bert](https://huggingface.co/UTibetNLP/tibetan_bert)), but does not include the training data or a detailed description of it. A Tibetan computer developer in Amdo, Sangjee Dondrub, has built a tokeniser for Tibetan using an improved version of BERT known as RoBERTa, available as of May 2022 at <https://huggingface.co/sangjeedondrub/tibetan-roberta-causal-base> (accessed January 29, 2025).

document classification,<sup>24</sup> sentiment analysis, question answering, NER, and text summarisation. It is not designed for non-classificatory tasks like text-generation or for those requiring logical reasoning or domain-specific knowledge. Its tokenisation method is based not on linguistic rules but on frequency, and it splits “words” or syllables into sub-tokens or “sub-words”, often consisting of a single character or character stack, which are not always recognisable to a human reader and so cannot be checked. A similar tokeniser is used by Tibetan LLaMA (Lv *et al.* 2024), a transformer-based LLM based not on BERT but on LLaMA,<sup>25</sup> the open-source LLM created by Meta (the owner of Facebook) as a competitor to OpenAI’s GPT-4 or Microsoft’s Bing Chat.<sup>26</sup> These tokenisation methods would not be suitable for a tool or procedure that needs to recognise actual words and to learn their linguistic functions.

Since late 2023, however, mainstream LLMs trained primarily on English texts have begun to demonstrate Tibetan-language capabilities. Initially, their ability was somewhat limited, probably because of the limited amount of raw training data in Tibetan fed to them by that stage. For example, in April 2024 we asked GPT-4 to

---

<sup>24</sup> When tested on a validation sample of news articles in Tibetan, Sun *et al.* (2022) reported that their TiBERT model correctly classified 86% of the unseen texts. Tibetan BERT also achieved an 86% accuracy rate on a similar test, classifying unseen texts in a partition of its news article training corpus, suggesting that Tibetan BERT has no inherent disadvantage over TiBERT despite its much lower computational demand.

<sup>25</sup> The newest version of LLaMA is at <https://www.llama.com> (accessed January 29, 2025).

<sup>26</sup> Tibetan Lama (T-LLaMA) was trained on 11 gigabytes of data in Tibetan and has so far been shown to be effective at text classification, basic news text generation and text summarisation. The developers note that it needs further development if it is to be ready for “tasks such as dialogue, reasoning, and translation” (Lv *et al.* 2024: 72). The T-LLaMA model is available at <https://huggingface.co/Pagewood/T-LLaMA>. It used a tokeniser called SentencePiece, which, like BERT, basically treats each character or character stack as a token and calculates probabilities for which token might follow it – again, not a suitable method for us to use to tokenise our training data for the spaCy language model.



tokenise a sample Tibetan paragraph taken from an online news article:<sup>27</sup>

གྲུང་གོ་ས་ཡོམ་ལས་ཁུངས་ ན་བས་དངོས་སུ་ཚད་འཇལ་གཏན་འཁེལ་བྱས་པར་གཞིགས་ན། ཟླ་དང་ལོའི་ཚེས་23ཉིན་གྱི་ཚུ་ཚོད་2དང་སྐར་  
མཉེན་ལྷན་ཅང་ཨ་ཁི་སུའུ་ས་ཁུལ་ལྷན་ཁྲུང་གི་ཚོང་དུ་རིམ་པ་7.1ཅན་གྱི་ས་ཡོམ་བྱུང་བ་དང་། xxཡོམ་ཁུངས་ཀྱི་གཏིང་ཚད་སྟེ་ལེ་22ཡིན་ལ། ཡོམ་  
ལྷན་གྱི་བྱང་གི་འབྲེད་ཐིག་ཏུ་41.26དང་ཤར་གྱི་གཞུང་ཐིག་ཏུ་78.63ཡིན་པ་ཤིང་།

GPT-4, although it was already more advanced in Tibetan than its predecessor GPT-3.5, responded by dividing the sentence into five utterances or phrases each ending with a *shad*, seeing each utterance as a token. It thus could not, at this stage, distinguish between a word and an utterance. By September 2024, it could identify individual words, but in most cases missed polysyllabic words (for example, it tokenised ས་ཡོམ་ལས་ཁུངས་ as four separate words). By that time, however, Google’s Gemini Ultra (Gemini Pro 1.5) model was already able to perform effectively when asked to tokenise the first two phrases of the above sentence, producing good results:

- གྲུང་གོ་ / ས་ཡོམ་ལས་ཁུངས་ / ན་བས་ / དངོས་སུ་ / ཚད་འཇལ་ / གཏན་འཁེལ་ / བྱས་པར་ / གཞིགས་ན།  
གྲུང་གོ་ (China) / ས་ཡོམ་ (earthquake) / ལས་ཁུངས་ (department/agency) / ན་བས་  
(strictly) / དངོས་སུ་ (in reality/actually) / ཚད་འཇལ་ (measurement) / གཏན་འཁེལ་  
(fix/establish) / བྱས་པར་ (having done) / གཞིགས་ན། (when looking  
at/considering)
- ཟླ་དང་ལོ་ / འི་ / ཚེས་23 / ཉིན་གྱི་ / ཚུ་ཚོད་2 / དང་ / སྐར་མཉེན་ / ལྷན་  
ཟླ་དང་ལོ་ (first month) / འི་ (of: genitive particle) / ཚེས་23 (date: 23rd) / ཉིན་གྱི་  
(of the day) / ཚུ་ཚོད་2 (hour: 2) / དང་ (and) / སྐར་མཉེན་ (minute: 9) / ལྷན་  
(on/upon)

On a separate test in September 2024, Gemini had improved enough to give results similar to those of Claude and Gemini Ultra. In that test, we compared their results to those given by the linguist and Tibetologist Camille Simon for the 40-syllable sentence བརྗོད་དོན་གཙོ་བོའི་སློབ་གསོ་  
ལག་གཉིས་པ་སློབ་འཇུག་ཞིབ་ཚགས་བྱས་ནས་ལག་བསྟར་བྱ་རྒྱུ་ནི་མིག་སྲེའི་བོད་སྲོང་ས་ཡོམ་ཁུངས་ཀྱི་ས་གནས་ལག་དང་ཚན་པ་ལག་གི་གཤམ་ཚེའི་ལས་དོན་  
ཞིག་ཡིན། (see Kyogoku *et al.* 2025 in this issue, Appendix F). ChatGPT4 and

<sup>27</sup> [http://tb.tibet.cn/tb/index/news/202401/t20240123\\_7562609.html](http://tb.tibet.cn/tb/index/news/202401/t20240123_7562609.html).

ChatGPT4o both tokenised this almost entirely on the *tsheg*, ignoring all two-syllable words except for two. Their error rate was 61.5% and 60% respectively, so we saw little improvement there. But Claude and Gemini had no problem in recognising most polysyllabic words and differed from the expert version only in not dividing particles, such as the genitive suffix, from a root word, which reflects a difference in tokenising method rather than an error. These two LLMs are thus already capable of tokenising a non-tokenised Tibetan passage based on their self-trained, probability-based understanding of the structure of the Tibetan language.

Nevertheless, although very promising overall, LLMs tend to be inconsistent in their decisions, so the same prompt may return slightly different results at different times. Gemini, when asked again to tokenise the sentence above about earthquakes three months later, divided it into 19 tokens rather than the 17 it had produced in its first attempt.<sup>28</sup> As they are fed more training data or their parameters are refined, the abilities of LLMs with a particular task or language – especially a low-resourced one – can even diminish (Pramanik 2025). In addition, with LLMs, there is no code or programming that an end-user can access or adjust apart from modification of their parameters.

Consequently, as Kyogoku, Erhard and their colleagues describe in their paper (section 3.1), the project chose to use Botok to tokenise the training data for the new language model. Botok has sufficient accuracy, it is consistent, it can easily be added to a pipeline or workflow so as to integrate with other tools such as spaCy, and Botok's open-source code allows the end-user to adjust their local versions should the need arise. Although Botok was designed to be used on classical Tibetan texts, the project team was thus able to adapt it for use with modern Tibetan texts by adding extensive wordlists of modern Tibetan terms to its internal dictionary (see Kyogoku *et al.* 2025, section 4.2). As a result, our adapted form of Botok, which we have termed

---

<sup>28</sup> In V1, Gemini rendered གཏན་འཁེལ་ and བྱམ་པར་ as two tokens, but treated them as one (གཏན་འཁེལ་བྱམ་པར་) in V2. It rendered དང་ལོ་ and འི་ as ལྷ་དང་ལོ་ and འི་ in V1 and as ལྷ་ and དང་ལོ་འི་ in V2. Numerals were not treated as separate tokens in V2, but grouped with their reference (i.e., ལྷ་ཚོ་2 / དང་ / ལྷ་མ་9) in V2.

“modern Botok”,<sup>29</sup> performs well with modern Tibetan, scoring 13/14 on a simple test (see Kyogoku *et al.* 2025, Appendix D).

## 6 Tagging in Tibetan

To develop a Tibetan language model for spaCy, we also needed to produce training data that is POS tagged, i.e., where each word in a text is identified according to its syntactic function as a noun, adjective, verb, or so forth. Annotation of this kind is also important for our project because it helps in the development of NER (named-entity recognition), a capability which we will need in order for our search tools to recognise automatically which words are names of places, people, offices and other entities.<sup>30</sup> In Section 3.2 of their paper, Kyogoku, Erhard and their colleagues describe their assessment of existing tools to see which might be suitable for producing POS tagged training data in modern Tibetan. They found that the rule-based tools were not able to make decisions based on context, and so could not resolve cases where a word might have more than one syntactic function. Neither could they adjust to unfamiliar genres or registers of writing. Shallow machine-learning tools such as ACTib produced much better results but, as noted above, had been trained on classical Tibetan and would again need to be fed large amounts of pre-annotated training data in modern Tibetan in order to work with modern texts.

As with tokenisation, however, the team found a remarkable improvement in transformer-based approaches to POS tagging. By

---

<sup>29</sup> See <https://github.com/Divergent-Discourses/modern-botok> and Erhard, Kyogoku *et al.* 2024. To compile this extended wordlist we combined relevant dictionaries from Christian Steinert’s collection ([https://github.com/christiansteinert/tibetan-dictionary/tree/master/\\_input/dictionaries/public](https://github.com/christiansteinert/tibetan-dictionary/tree/master/_input/dictionaries/public)) with Botok’s built-in Grand Monlam Dictionary and our own list of personal and place names derived from the newspaper corpus.

<sup>30</sup> Unlike those European languages which have capital letters and definite articles to mark some types of named entities, Tibetan offers few transparent tricks to identify them. The simplest heuristic, given the tendencies of Tibetan grammar, would be to search for proper nouns and their immediate neighbours.

September 2024, GPT-4 could POS-tag a Tibetan sentence with around 80-90% accuracy, and a comparison of POS tagging in Tibetan by GPT-4, Claude 3.5 Sonnet, and Gemini shows a similar error rate of 10-20% (see Kyogoku *et al.* 2025, Appendix F), which is low given the likely rate of progress.

The tagging by LLMs again showed inconsistencies, but, unlike with tokenisation, the resulting problems were not critical and in most cases could be corrected: by using rule-based coding at the post-processing stage, we were able to correct many of the recurrent errors in Gemini's tags. By August 2024, the Diverge team had therefore abandoned our initial plan to use a shallow machine-learning tool to produce POS tagged training data for our Tibetan language model and instead used Gemini Pro 1.5 to POS tag our training data. By late August, we had generated sufficient quantities of tokenised and POS tagged training data in modern Tibetan to develop from scratch the first Tibetan-language model for spaCy (Kyogoku *et al.* 2025, Sections 4 and 5).<sup>31</sup>

## 7 *Transformer-based Tools: Topic Modelling and Semantic Searching*

Our success in developing a Tibetan-language model for spaCy meant that, two years after we began the project, we were finally able to test our text-mining platform, the iLCM, for its capacity to read and analyse Tibetan texts. We have not yet completed the testing process, but initial results indicate that iLCM can process Tibetan without significant errors, although some issues remain concerning its handling of unfamiliar Unicode characters. We anticipate that other such problems will arise, as is frequently the case when using software developed for use with high-resource languages to work with low-resource languages. Overall, however, it is now likely that in time we will be able to use the full capabilities of the iLCM as well as related tools and platforms. This will then provide us with a sophisticated set

---

<sup>31</sup> For the Modern Tibetan spaCy code and data set, see Kyogoku *et al.* 2024b, c.

of tools to identify “divergent discourses” – changing topics or narrative foci – in our corpus.

However, we had to allow from the outset for the possibility that we would not be able to assemble sufficient annotated training data from modern Tibetan texts, or that, even if could, the iLCM and similar tools might not work with our Tibetan model. We therefore pursued a parallel strategy in our project which avoided dependency on tools requiring pre-annotated training data. This strategy, initiated and led by Ronald Schwartz, involved the development of tools that employ vector embeddings derived from transformer-based LLMs to analyse Tibetan texts. As we have seen with LLMs, these tools do not need tokenising or POS-tagging to have been carried out on their training data.

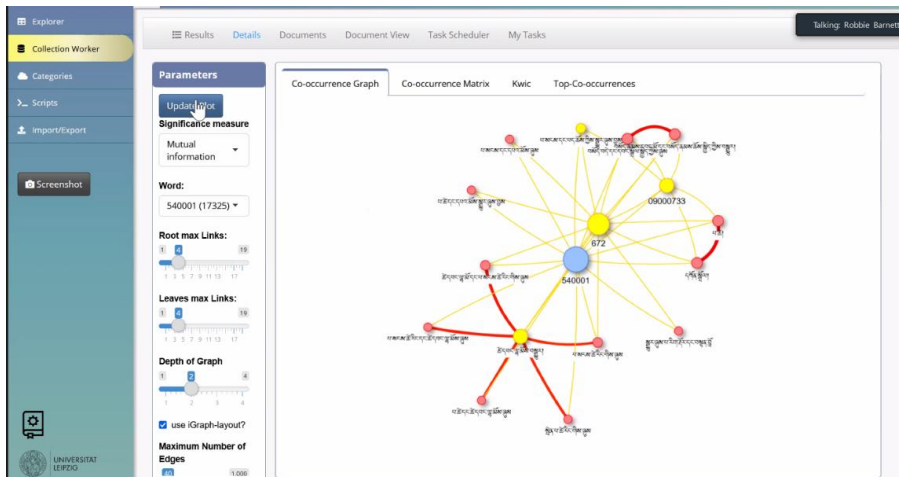


Figure 3 An iLCM visualisation showing relations between certain terms and names in the corpus

As **Ronald Schwartz** and **Robert Barnett** explain in their paper for this issue, “Religious Policy in the TAR, 2014–24: Topic Modelling a Tibetan-Language Corpus with BERTopic” (Schwartz & Barnett 2025), embeddings are numerical strings into which a word, phrase, sentence, or paragraph has been encoded. Those encodings are calculated with reference to context, so that the number used for a particular word or text is adjusted to account for the numbers of words or passages that adjoin it or are near it. This ability to incorporate

contextual information goes further than just word co-occurrences. Thus the encoding for the word “bear” in the sense of a large furry animal would differ from the same word used as a verb in the sense of carrying something. In addition, the word “bear” followed by the word “arms” would be encoded differently from the word “bear” followed by “fruit”. As a result, encodings of this kind do not just represent the lexical form of a word or text; they in effect convey the semantic content or meanings of those texts.

In the transformer-based approach, these encoded numbers are treated mathematically as if they are situated within a conceptual space that consists of multiple dimensions. As a result, numbers that are close together algebraically in the vector space represent words or texts that are close together in meaning. This has a consequence of great importance for NLP: it means that these transformer-based forms of vectorisation can carry out tasks that identify meanings of different kinds within a text, even if not apparent to a reader or a conventional search tool. As noted above, these transformer-based tools can perform these tasks without needing task-specific training data.

The project therefore set out to develop two such tools for use with modern Tibetan. One is topic modelling – the automatic, corpus-scale identification of textual foci. Topic modelling identifies topics or themes in a text or corpus based on the above-mentioned algebraic forms of detection, whether or not the explicit name of a given topic is mentioned. Topic modelling in this case is not guided – the tool determines for itself what are the different topics in a text. The second tool is semantic search, which is a guided form of the same technology: a user enters a query and the semantic search tool finds any texts in the corpus which resemble that query – even if those texts do not include any of the words found in the query. Thus, if one asked a semantic search tool to search for similar passages to the phrase “people who study Tibetan texts”, it could return not only all the passages in that corpus about Tibetanists, but also those about Buddhas, animals or deities that read novels in Sanskrit, Chinese or perhaps Uyghur. Both these tools are of particular value to textual analysts because by their nature they are likely to find passages or texts that are not anticipated by the searcher yet are relevant to the query.

At least in principle, they thus can contribute to some extent to decreasing the risk of confirmation bias by an analyst.

The paper by Schwartz and Barnett explains the principles of this approach to topic modelling and to semantic searching and demonstrates an application of the topic modelling tool in the Tibetan context. Using a set of some 4,000 “born-digital” articles collected by Schwartz from online Tibetan-language newspapers over the last ten years, the paper shows how Schwartz first cleaned up the texts, subdivided them into paragraphs or “chunks”, and compiled them into a corpus. He then created a “subcorpus” consisting only of articles including the Tibetan terms for “religion” or “Tibetan Buddhism”. The topic modelling tool was then run on the subcorpus to identify variations in discussions of religion in the subcorpus. It divided the paragraphs into groups, each with a different narrative emphasis. Schwartz then used an LLM called Claude Sonnet 3.5, which is already able to read and process Tibetan texts, to analyse the top ten paragraphs from each group (meaning those identified by the tool as most “representative” of the narrative focus of that group) and to add a label and a set of keywords that typified the topic of that group. With this information, Schwartz and Barnett were then able to identify a number of topics (about half of the total number of topics) indicating drives by the Chinese authorities to regulate the behaviour of monks and nuns in Tibetan monasteries. They also identified a third of the topics as representing ideological drives to reshape thinking about acceptable forms of religious belief among the Tibetan public, and a set of topics or themes that represented ongoing background arguments or opinions circulated by the Chinese government concerning religion in Tibet. Since the topic model could display its findings in terms of occurrences of each topic over time (a function known as “dynamic topic modelling”), the analysts could also map the duration of each drive and identify their peaks of activity, in terms of occurrences of the topics in the official Tibetan media.

In their paper for this issue, “Developing a Semantic Search Engine for Modern Tibetan” (Engels & Barnett 2025), **James Engels** and **Robert Barnett** describe Engels’ construction of a semantic search engine for Tibetan texts that can be used by members of the public

without any knowledge of coding or other computational techniques. They first explain the differences between semantic searches and conventional keyword searching. In the latter, a user finds all words in a corpus that are identical to the query term or (in the case of fuzzy searches or lemmatized systems) resemble it lexically; in the former, search results produce texts similar in meaning, but not necessarily in form, to the query. The paper goes on to explain the concepts behind the historical development of semantic searching, describing the use of transformer-based approaches to vectorisation and summarising in lay terms the broad theories underlying such a system and the history of these major advances in NLP. In particular, they credit Schwartz with realizing that one company specialising in AI, Cohere, has already developed the technology for creating vector embeddings for less-resourced languages, including Tibetan.<sup>32</sup> In effect, Cohere's embedding model can "understand" Tibetan at an advanced level of complexity. Cohere allows any user to submit a text or an entire corpus which it then converts into embeddings and returns to the user, for a small fee. It is this that makes semantic searching possible in Tibetan. In addition, Cohere's model is multilingual: since it deals with meanings numerically, translation is an emergent feature of its capabilities. With Cohere's embeddings, one can thus submit a query in Chinese or Vertical Mongolian and get results in Tibetan or whatever is the language of one's corpus, if Cohere has been trained on that language or one that is linguistically similar.

The paper also provides more detail about the distinction between topic modelling and semantic searching. Engels characterises the former as "outside-in" analysis of a corpus, such as looking at what topics tend to occur over time within a corpus or identifying what is represented in a particular article. He describes semantic searching as an "inside-out" method, where a researcher knows what kinds of things they want or expect to see but do not know where to find them or what words might be used to signify them.

---

<sup>32</sup> Cohere's Multilingual Model 2.0 is at <https://cohere.com/> (accessed January 15, 2025). Both Schwartz and Engels advise against using Cohere's Multilingual Model 3.0 for Tibetan texts.



In the second half of his paper, Engels describes the techniques he used to create a semantic search tool for Tibetan that would not require programming knowledge by its users but would be fully accessible to the public. By adapting code initially developed by Schwartz for semantic searching first in Chinese and then in Tibetan, Engels produced a publicly accessible version of the tool for use with Tibetan texts that is hosted on a university server and has a simple user interface.<sup>33</sup> It lists the date, source, title of article for each search result and allows the user to rank results either by relevance (semantic match) or by date, or by one of the other metadata fields. It includes a toggle allowing non-readers of Tibetan to see a translation of the query or the results in English. For this he linked the system to Bing Translate, which currently appears to be the most reliable provider of online translations from Tibetan to English.

Currently, the projects' semantic search engine is linked to a test corpus of recent Tibetan newspaper articles harvested from online sites, but in future it will be linked to the Divergent Discourses corpus of historical newspapers. It is, we believe, the first online system for searching historical Tibetan newspapers using state-of-the-art NLP tools. It can easily be linked to any corpus of Tibetan (or other) texts by anyone with basic programming knowledge; search systems of this kind, Engels notes, are not difficult to build and do not require expertise in software design.

### 8 *Identifying a discourse: 'Friendship' in the Tibet Mirror*

The final paper in this issue is **Natalia Mikhailova's** study, "*The Tibet Mirror, 'Friends' of Tibet, and the Internationalisation of the Tibet Question*" (Mikhailova 2025). Her article offers an example of the project's aim – the identification and discussion of discourses in Tibetan-language newspapers from the 1950s to the early 1960s. Her contribution is distinct from the other papers in this issue, firstly in

---

<sup>33</sup> The public link to the project's semantic search tool is <https://tibetcorpus.uni-leipzig.de/search/> (accessed January 30, 2025).

that it is an example of applied research rather than a discussion of methodology, and secondly, in that it is an example of research in the pre-digital humanities era: she carried out the study without the advantage of text-mining tools, before our corpus has become available, and before our tools have been completed. Using nothing more than photographs or scans of newspaper pages obtained from libraries, and studying the often imperfect images without the aid of any search or other tools, Mikhailova shows the enormous labour that goes into research of this kind, identifying themes in the articles, supplying lengthy translations, and connecting them to their historical context and, where available, to previous studies.

Her focus is on the most prominent of all Tibetan-language newspapers outside Tibet, the *Tibet Mirror* (the *Yul phyogs so so'i gsar 'gyur me long*). The *Mirror* was published from the north-eastern Indian city of Kalinpong from 1925 to 1963, and Mikhailova's study looks at its output in the years following the annexation and incorporation of Tibet by the People's Republic of China (PRC) in the early 1950s. She includes the years after the mass exodus of Tibetans and their leader, the 14<sup>th</sup> Dalai Lama, from Tibet to India following the failed uprising of March 1959. Her main finding is that, among other discourses, the *Tibet Mirror* aimed to construct for its readers a narrative of international support for Tibet in its struggle against the PRC. The paper, she shows, focused at times on presenting the governments of Great Britain, India, and the United States as "friends" or supporters of the Tibetans, particularly those who had fled into Exile. The paper also represented the Chinese nationalist government in Taiwan as supportive of the exiles, and even as having said that, if it were to regain power in China, it would support Tibet's independence. She also shows how the paper proposed arguments, based on Nepal's application to the United Nations for membership of that body in 1949, that could be used to argue for Tibet's independence. As she explains, these arguments came to dominate exile discussions of Tibetan independence for many decades afterwards.

This study of a single narrative thread in one Tibetan newspaper offers an early example of the kind of findings that are of interest to the project, and which other team members might explore in future

studies. Those studies, however, will have the advantage of access to the project’s computational tools and to its corpus, which will shortly be ready for use. Over the next year, as the project moves from the development of text-mining tools to the application of those tools to the study of discourses in Tibetan newspapers, we will see how the use of digital methodologies confirms, extends or varies Mikhailova’s findings.

## 9 Conclusion

At the current rate of progress, a transformer-based Tibetan LLM will soon outmode even the best tools based on earlier approaches. At the time of writing, Monlam AI’s promised first-generation Tibetan LLM (including various functions like a Dalai Lama chatbot), which is being developed jointly with Esukhia, is still in early stages of development, but it is likely to expand its abilities rapidly as its range of training data is expanded.<sup>34</sup> At Berkeley, Sebastian Nehrdich and colleagues have produced the initial elements of a Tibetan transformer model.<sup>35</sup> If Tibetan LLM technology improves enough to gain “intuitive” understandings of modern Tibetan, it is possible that dedicated POS taggers will no longer be needed. Meanwhile, a number of English-language-based LLMs from major corporate entities have advanced astonishingly quickly in their capacity to understand and analyse Tibetan text, albeit with different strengths. At the end of 2023, no public-facing LLM could produce intelligible interpretations of Tibetan data. By the autumn of the following year, some major LLMs – such as OpenAI’s GPT-4, Google’s Gemini, and Anthropic’s Claude 3.5 Sonnet – had developed translation capacity and other

---

<sup>34</sup> See <https://monlam.ai/about> (accessed December 18, 2024), and <https://github.com/MonlamAI> (accessed December 18, 2024).

<sup>35</sup> See <https://huggingface.co/buddhist-nlp/byt5-mitra-bo>. Other initiatives include “TibetaMind” (“an advanced language model based on the Llama 3-8B-Instruct architecture, further fine-tuned using extensive Tibetan language corpora”, available at <https://huggingface.co/DaydreamerF/TibetaMind>) and the T-LLaMA model (Lv *et al.* 2024). Both sites accessed January 30, 2025.

competencies in modern Tibetan. For example, we tested them with the following sentence:

གྲུང་གོང་དོ་ཤོ་ལ་བྱེད་མཁན་རྒྱལ་ཕྱོགས་ཕྱོགས་གཏོགས་ཀྱིས་དྲ་ལའི་རུ་ཚོགས་ལ་བརྟེན་ནས་གྲུང་གོ་ལ་ཕྱལ་འཛོལ་ཞིག་ཏུ་གཏོང་བའི་ལྷོག་གཡོ་གཤོས་རྒྱ་སྐབར་ཕྱི་གསལ་དུ་འགྱུར་བ་བྱུང་མེད་ལ། དྲ་ལའི་རུ་ཚོགས་ཀྱིས་“བོད་རང་བཙན”ཡོང་འབྱེད་པའི་ལ་ཕྱལ་འཛོལ་ལུགས་ཀྱི་འོ་ཡང་སྐབར་ཕྱི་གསལ་དུ་འགྱུར་བ་བྱུང་མེད།

GPT-4’s translation of this sentence, as of September 2024, was “Western-backed anti-China forces have not yet resorted to external interference or attempted to secretly incite a sudden breakup of China through reliance on groups like the Dalai clique.”<sup>36</sup> This reversed the actual meaning of the sentence and missed a number of details.<sup>37</sup> Claude, however, translated it correctly as “The Western forces opposing China have not changed their covert plot to use the Dalai Lama's group to split and disintegrate China, whether in the past, present, or future. Similarly, the separatist nature of the Dalai Lama's group's hope for "Tibetan independence" has also not changed over time.”<sup>38</sup> Gemini’s translation was similar.<sup>39</sup> The translations by both

<sup>36</sup> <https://chatgpt.com/c/66dae400-2168-8002-975c-e082e71bd3c7>. No longer accessible.

<sup>37</sup> GPT-3.5 was almost entirely incorrect except for a few words, underlined here: “Those who engage in the struggle for the establishment of Tibet's independence, from the perspective of the Tibetan people, are not caught in the virtue of a particular kind of nobility. With the justification of the Tibetan struggle becoming apparent, there is no change in the recognition of the essence and nature of the pursuit of 'Bod rang btsan' (Tibetan independence) even three times in the past” (<https://chat.chatbotapp.ai/chats/-O6sEh0wWsm0uJUb0fXP?model=gpt-3.5>; no longer accessible).

<sup>38</sup> See <https://claude.ai/chat/6faa50e1-f10b-44f7-98c0-ce854b6037a1> (accessed January 30, 2025).

<sup>39</sup> “The Western powers, who oppose China, have relied on the Dalai Lama group to split China. However, their covert attempts to achieve this have not changed from the beginning to the end. Similarly, the Dalai Lama group's separatist ideology, which seeks "Tibetan independence," has not changed from the beginning to the end” (<https://gemini.google.com/app/eaaf83e2455372f9>; accessed January 30, 2025).

Claude and Gemini were far more precise and detailed than the online translations of the same sentence by Bing or Google.<sup>40</sup>

Claude Sonnet has demonstrated even more advanced Tibetan-language capability. It can produce high-quality translations of modern texts, perform generative tasks such as summarising, finding keywords, and generating labels, and it can do these directly in Tibetan without working through an English translation. GPT-4o, unlike previous generations of GPT, also is capable of summarising and extracting themes and topics from modern Tibetan texts in a remarkably human-like way. Future generations of LLMs are thus likely to have the capacity to understand and analyse Tibetan texts to a degree not markedly different from any other language for which the machine has been supplied with sufficient training data.

For those working with Tibetan texts, this means that in the near future, transformers will likely be the tool of choice for translation, summarisation, classification, topic modelling and semantic searches. They will also be very strong options for tokenising and POS tagging, if those tasks are still needed. However, they remain “black-box” technologies: the end-user cannot adjust their code or even be certain as to how they reach their conclusions, and their results and capabilities are likely to vary over time. Given these uncertainties, the Tibetan studies community is probably going to continue for some time to need to maintain and develop tools based on more traditional approaches, whether based on rules or machine learning, in order to have robust and replicable techniques available for the analysis of Tibetan texts.

Overall, however, we have found that for our project some objectives are most easily achieved by using more traditional

---

<sup>40</sup> Google Translate, which had only recently made Tibetan available, struggled to translate this sentence and rendered it as “Western opponents of China rely on the Dalai Lama's party to see China as a divided country There is no change in the smuggling [*sic*]. The Dalai Lama's group's ‘Tibetan independence’ is also a form of separatism. It hasn't changed.” Bing Translate was more accurate than Google Translate but also missed some details, rendering this as “The conspiracy of the Western forces opposing China to use the Dalai clique to split China has never changed. The separatist nature of ‘Tibetan independence’ has never changed.”

approaches, not least because, firstly, these approaches often come with pre-developed user interfaces and, secondly, their underlying code can be re-engineered if needed. Their results are, in addition, more consistent. Other tasks, such as semantic searching, are better served by transformer-based techniques. As a result, we have found that, so far, a combination of traditional and more recent methods is preferable.

This survey of the computational tools and strategies developed or assessed for the Divergent Discourses project will, we hope, be useful for Tibetanists looking to work with digital methodologies such as text-mining. However, our survey is by no means comprehensive, focusing mainly on tools developed by BDRC, Esukhia, and ACTib, and we apologise for important contributions that we have not included here. We note, however, that most of the steps that we have described here involve specialist technical knowledge and experience, and to put them into practice takes extensive time, funding, and expert support for problem-solving. Implementation of these tools is not smooth and involves multiple, time-consuming stages of error correction, troubleshooting and consultation within the specialist community. Nevertheless, we hope this special issue of RET, by describing some of the technical considerations we have experienced in developing text-mining tools for use with modern Tibetan texts, will be helpful for others seeking to apply such methods in their work.

### Bibliography

Engels, James, and Robert Barnett

“Developing a Semantic Search Engine for Modern Tibetan,” *Revue d'Etudes Tibétaines* 74, 2025, pp. 262–283.

Engels, James, Robert Barnett, Franz Xaver Erhard, and Nathan. W. Hill  
 “Transkribus\_utils: Paragraph Extractor (v1\_Paragraph\_Extractor),” *Zenodo*, 2024. [doi:10.5281/zenodo.10810509](https://doi.org/10.5281/zenodo.10810509)

- Engels, James, Franz Xaver Erhard, Robert Barnett, and Nathan W Hill  
 “Tibetan for Spacy 1.1 [Data set],” *Zenodo*, 2023. [doi:10.5281/zenodo.10148636](https://doi.org/10.5281/zenodo.10148636)
- Erhard, Franz Xaver.  
 “The Divergent Discourses Corpus: A Digital Collection of Early Tibetan Newspapers of the 1950s and 1960s,” *Revue d’Etudes Tibétaines* (74), 2025a, pp. 45–81.
- “Text and Layout Recognition for Tibetan Newspapers with Transkribus,” *Revue d’Etudes Tibétaines* (74), 2025b, pp. 129–172.
- Erhard, Franz Xaver and Xiaoying 笑影  
 “Foreign names and places in Tibetan newspapers of the 1950s and 1960s,” *Revue d’Etudes Tibétaines* (74), 2025, pp. 173–187.
- Erhard, Franz Xaver, Yuki Kyogoku, Robert Barnett, and Nathan W. Hill  
 “Modern-Botok. Custom dictionary for modern Tibetan (v0.1) [Data set],” *Zenodo*, 2024. [doi:10.5281/zenodo.14034747](https://doi.org/10.5281/zenodo.14034747).
- Erhard, Franz Xaver, Xiaoying 笑影, Barnett, Robert, and Nathan W. Hill  
 “Tibetan Modern U-chen Print (TMUP) 0.1: Training Data for a Transkribus HTR Model for Modern Tibetan Printed Texts,” *Fachinformationsdienst (FID) Asien*, 2023. [doi:10.48796/20240313-000](https://doi.org/10.48796/20240313-000).
- “Toponyms and Anthroponyms from Tibetan-language Newspapers of the 1950s and 1960s: Three Name Lists,” *Zenodo*, 2024. [doi:10.5281/zenodo.14526125](https://doi.org/10.5281/zenodo.14526125).
- Dakpa, Jamyang, Tashi Dhondup, Yeshe Jigme Gangne, Edward Garrett, Marieke Meelen, and Sonam Wangyal  
 “Modern Tibetan Corpus,” *Github repository*, 2021a. Available online at <https://github.com/tibetan-nlp/modern-tibetan-corpus/tree/v1.0> (accessed January 30, 2025).

"Modern Tibetan Corpus Annotated for Verb-argument Dependency Relations," *Zenodo*, 2021b, [doi:10.5281/zenodo.4727129](https://doi.org/10.5281/zenodo.4727129).

Faggionato, Christian, Edward Garrett, Nathan W. Hill, Samyo Rode, Nikolai Solmsdorf, and Sonam Wangyal

"Classical Tibetan Corpus Annotated for Verb-argument Dependency Relations," *Zenodo*, 2021. [doi:10.5281/zenodo.4727108](https://doi.org/10.5281/zenodo.4727108).

Faggionato, Christian, Nathan W. Hill, and Marieke Meelen

"NLP Pipeline for Annotating (Endangered) Tibetan and Newar Varieties." In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pp. 1–6, Marseille. European Language Resources Association. 2022. Online available at <https://aclanthology.org/2022.euralli-1.1/> (accessed January 31, 2025).

Hill, Nathan W., and Edward Garrett

"A part-of-speech (POS) tagged corpus of Classical Tibetan [Data set]," *Zenodo*, 2017. [doi:10.5281/zenodo.574878](https://doi.org/10.5281/zenodo.574878).

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: ICDAR 2017: proceedings*. Los Alamitos: Conference Publishing Services, IEEE Computer Society, 2017, pp. 19–24. [doi:10.1109/ICDAR.2017.307](https://doi.org/10.1109/ICDAR.2017.307)

Kyogoku, Yuki, Franz Xaver Erhard, Robert Barnett, and Nathan W. Hill

"TibNorm - Normaliser for Tibetan (Version v1)," *Zenodo*, 2024a, [doi:10.5281/zenodo.10815272](https://doi.org/10.5281/zenodo.10815272).

"Diverge-Gemini POS-tagged Corpus of Modern Tibetan (1.0) [Data set]," *Zenodo*, 2024b, [doi:10.5281/zenodo.14447192](https://doi.org/10.5281/zenodo.14447192).



"Basic Modern Tibetan SpaCy Model," *Zenodo*, 2024c, [doi:10.5281/zenodo.14494472](https://doi.org/10.5281/zenodo.14494472).

Kyogoku, Yuki, Franz Xaver Erhard, James Engels, and Robert Barnett  
"Leveraging Large Language Models in Low-resourced Language NLP: A spaCy Implementation for Modern Tibetan," *Revue d'Etudes Tibétaines* (74), 2025, pp. 188–221.

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](https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_72_02.pdf) (accessed January 20, 2025).

Lv, Hui, Pu Chi, La Duo, Yan Li, Zhou Qingguo and Shen Jun  
"T-LLaMA: a Tibetan large language model based on LLaMA2," *Complex & Intelligent Systems* 11(1), 2024. [doi:10.1007/s40747-024-01641-7](https://doi.org/10.1007/s40747-024-01641-7).

Marieke Meelen, Nathan W. Hill, and Christopher Handy  
"The Annotated Corpus of Classical Tibetan (ACTib), Part I—Segmented version, based on the BDRC digitised text collection, tagged with the Memory-based Tagger from TiMBL," *Zenodo*, 6 July 2017a. [doi:10.5281/zenodo.823707](https://doi.org/10.5281/zenodo.823707).

"The Annotated Corpus of Classical Tibetan (ACTib), Part II—POS-tagged version, based on the BDRC digitised text collection, tagged with the Memory-based Tagger from TiMBL," *Zenodo*, 2017b. [doi:10.5281/zenodo.822537](https://doi.org/10.5281/zenodo.822537).

Meelen, Marieke, Sebastian Nehrlich and Kurt Keutzer  
"Breakthroughs in Tibetan NLP & Digital Humanities," *Revue d'Etudes Tibétaines*, (72), 2024, pp. 5-25. Available online at [https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret\\_72\\_01.pdf](https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_72_01.pdf) (accessed January 26, 2025).

Meelen, Marieke, and Élie Roux

"The Annotated Corpus of Classical Tibetan (actib) - Version 2.0 (segmented & Pos-tagged)," *Zenodo*, 2020. [doi:10.5281/zenodo.3951503](https://doi.org/10.5281/zenodo.3951503).

Meelen, Marieke, Élie Roux, and Nathan Hill

"Optimisation of the Largest Annotated Tibetan Corpus Combining Rule-Based, Memory-Based, and Deep-Learning Methods." In *ACM Transactions on Asian and Low-Resource Language Information Processing* 20 (1), 2021, pp. 1–11. [doi:10.1145/3409488](https://doi.org/10.1145/3409488).

Natalia Mikhailova

"*The Tibet Mirror*, 'Friends of Tibet,' and Internationalisation of the Tibet Question," *Revue d'Etudes Tibétaines*, (74), 2025, pp. 284–328.

Pramanik, Siddhartha.

"Continual Learning and Catastrophic Forgetting: The Challenges and Strategies in AI," *Medium*, January 17, 2025. Available online at <https://medium.com/@siddharthapramanik771/continual-learning-and-catastrophic-forgetting-the-challenges-and-strategies-in-ai-636e79a6a449> (accessed January 30, 2025).

Sabbagh, Christina.

"Enhanced HTR Accuracy for Tibetan Historical Texts - Optimising Image Pre-processing for Improved Transcription Quality," *Revue d'Etudes Tibétaines*, (74), 2025, pp. 82–128.

Sabbagh, Christina, Franz Xaver Erhard, Robert Barnett, and Nathan W. Hill

"Divergent Discourses Custom Image Preprocessing (Sauvola Binarisation)," *Zenodo*, 2024a. [doi:10.5281/zenodo.14525692](https://doi.org/10.5281/zenodo.14525692).

"Divergent Discourses Custom Image Preprocessing (Forked Binarisation)," *Zenodo*, 2024b. [doi:10.5281/zenodo.14523007](https://doi.org/10.5281/zenodo.14523007).

Schwartz, Ronald, and Robert Barnett

“Religious Policy in the TAR, 2014–24: Topic Modelling a Tibetan Language Corpus with BERTopic,” *Revue d’Etudes Tibétaines*, (74), 2025, pp. 222–261.

Sun Yuan, Liu Sisi, Deng Junjie, Sun Yuan and Zhao Xiaobing

“TiBERT: Tibetan Pre-trained Language Model\*.” *In* IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2022, pp. 2956–2961. [doi:10.48550/arXiv.2205.07303](https://doi.org/10.48550/arXiv.2205.07303).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin

“Attention is All you Need,” *Advances in Neural Information Processing Systems*. 30. Curran Associates, 2017. [doi:10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762)

Wallman, Jeff, Zach Rowinski, Ngawang Trinley, Chris Tomlinson, and Kurt Keutzer

“Collection of Tibetan Etexts Compiled by the Buddhist Digital Resource Center,” *Zenodo*, 2017. [doi:10.5281/zenodo.821218](https://doi.org/10.5281/zenodo.821218).

Zhang Jiangyan, Deji Kazhuo, Luosang Gadeng, Nyima Trashi, and Nuo Qun


“Research and Application of Tibetan Pre-training Language Model Based on BERT.” *In* Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics (ICCIR '22). Association for Computing Machinery: New York, 2022, pp. 519–524. [doi:10.1145/3548608.3559255](https://doi.org/10.1145/3548608.3559255).



# **The Divergent Discourses Corpus: A Digital Collection of Early Tibetan Newspapers from the 1950s and 1960s**

Franz Xaver Erhard

(Leipzig University)

he production of digital corpora is standard procedure for libraries. However, because of the relatively high cost, digitising archival holdings everywhere is still not a common practice, particularly beyond the global north. For Tibetan studies, it is a sad reality that the plethora of invaluable material is squirreled away in too often inaccessible archives in the People's Republic of China (PRC), uncatalogued boxes in Western libraries, or unexplored private collections. In the case of newspapers, the situation is even more problematic. Neither Chinese nor Indian institutions have systematically collected and preserved Tibetan language newsprint, or they may be reluctant to make it accessible to the public. The same is probably true for the early newspaper publishing houses and editorial offices that often lack the interest or funds to sustain a continuous publication archive. Moreover, many offices ceased to exist, and rapid and dramatic economic development has led the sources to disappear without a trace. The value of newspapers as a detailed historical record is too often underestimated; or, to the contrary, e.g., in the People's Republic, newspapers are seen as containing sensitive information that needs to be censored and controlled.

The Divergent Discourses project<sup>1</sup> studies the role of narrative and discourse in the perpetuation of antagonisms in the Tibet-China dispute in the formative period of the 1950s and 1960s. For studying Tibetan history and societal change, Tibetan language newspaper archives can constitute a significant source.<sup>2</sup> They document contemporary events and debates in minute detail and reflect the slow and gradual changes in language, society, and topics that would otherwise escape scholarly attention. To facilitate the study of Tibetan language discourses in the communities within the PRC and in South Asia, the project compiled a corpus of Tibetan language newspapers, the Divergent Discourses Corpus (DD Corpus).

At the time of writing, the DD Corpus contains 16,718 pages from 16 newspapers published between 1950 and 1965 in India and the PRC. The newspapers have been sourced from the holdings of universities, libraries, and archives outside the PRC, accumulating a substantial number of Tibetan-language newspapers from that period. The corpus does not, however, contain a complete set of any one newspaper and cannot claim to represent all newspapers published in Tibet and Tibetan areas. There were probably more newspapers, especially on the prefectural and county levels, whose names and histories have not survived the turbulent years of the 1950s to 1970s in the PRC.

Most of the newspapers in the corpus were scanned by the project,<sup>3</sup> and will – where copyright permits – become openly available online

---

<sup>1</sup> The Divergent Discourses project received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (<https://gepris.dfg.de/gepris/projekt/508232945?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/> (accessed on January 10, 2025) and the other contributions to this special issue.

<sup>2</sup> When using the term Tibet, I do not refer to a Tibetan country, or Tibetan region, etc. but imply the broadest sense of the term, encompassing all Tibetan communities in the PRC and beyond, including in India and Nepal.

<sup>3</sup> Where available, we relied on copies previously scanned by other institutions, such as the Columbia University Libraries (CU), the Oriental Institute of the Czech Academy of Sciences (OI), and the Library of Tibetan Works and Archives (LT). We would like to thank the Leipzig University Library and the Staatsbibliothek zu Berlin for their support in digitising materials for the project.

at Staatsbibliothek zu Berlin's Crossasia repository starting in 2025.<sup>4</sup> Besides the digital images, the newspaper corpus will be made available as machine-readable e-text to facilitate studies using Digital Humanities approaches and tools.<sup>5</sup>

The following sections will describe the DD Corpus and its context of the 1950s and 1960s to provide a more comprehensive picture of the material it contains, as well as to point out the gaps and limitations researchers should be aware of when using the material.

### 1 *Publication and Production of Tibetan Newspapers*

Tibetan newspapers are a relatively recent phenomenon.<sup>6</sup> The concept and technology were introduced to the Tibetan cultural sphere only in 1904 by the German missionary August Hermann Francke (1870–1930)<sup>7</sup> and independently by the Chinese Ambans Lian Yu (聯豫, born 1886, office in Lhasa 1906–1912) and Zhang Yintang (張蔭棠, 1860–1935, office in Lhasa 1906–1907) in 1907.<sup>8</sup> Before 1950, only a handful of Tibetan newspapers – produced mainly by Moravian missionaries – existed on the southern slopes of the Himalayas, including the well-known *Tibet Mirror* (TIM, founded 1925). When the Chinese People's Liberation Army (PLA) annexed Tibet to the newly established PRC, some Tibetan language newspapers were published in the Republic of China, e.g. the bilingual 'Vernacular News in Tibetan Language' (*Bod*

---

<sup>4</sup> <https://crossasia.org/en/> (accessed on January 10, 2025).

<sup>5</sup> For a technical and more systematic corpus description, see the appendix.

<sup>6</sup> Tibetan newspapers and their history are with a few notable exceptions mostly unstudied. A general overview of newspaper publishing in the first half of the 20<sup>th</sup> century was provided in English e.g. by Sawerthal 2018 and Erhard *et al.* 2018; in Chinese by Xu 2003, Zhou 2005; in Tibetan Klu ma tshal 2001, 2009. For the second half of the 20<sup>th</sup> century, see e.g. Shar ba thog med 1999, Zhou 2005, Hartley 2005, and Erhard 2015.

<sup>7</sup> On Francke's *Ladakh Akbar*, see Erhard 2021: 272–279; Erhard & Hou 2018: 4–7; Römer & Erhard 2007: 242–247; Walravens & Engelhardt 2010; and Walravens 2002.

<sup>8</sup> On the ambans see Ho 2008; Kobayashi 2020; on their newspaper, see Erhard & Hou 2018: 8–9.

*yig phal skad kyi gsar 'gyur*) by the Mongolian and Tibetan Affairs Bureau (MTAB) in Beijing.<sup>9</sup>

In the societies of the global north, newspapers are attributed the democratic role of checking the reach of government and its institutions and, at the same time, providing a forum for public exchange and debate, thereby creating the public sphere, which in the words of Jürgen Habermas, is a sphere “which mediates between society and state, in which the public organizes itself as the bearer or public opinion, [and] accords with the principle of the public sphere – that principle of public information which once had to be fought for against the arcane policies of monarchies and which since that time has made possible the democratic control of state activities” (Habermas 1974: 50). As such the concept of “public sphere is a ‘political term’ that relates to the democratic system” (Fiedler & Meyen 2015: 837).

In less democratic or even autocratic societies, such as Tibet and the People’s Republic of China, neither a democratic public sphere nor a ‘free press’ existed. Here, newspapers were published and controlled by the state or state-like institutions.<sup>10</sup> Nevertheless, the state’s or government’s “leaders ... despite all pretence to the contrary (‘The Party is always right’, ‘Dictatorship of the Proletariat’), could not simply represent power, they had to legitimize it through public communication” (Fiedler & Meyen 2015: 837). Consequently, authoritarian governments use their force to build a confined, authoritarian public sphere by “limiting the range of topics that can be discussed openly.”

On the other hand, the state wants to create the appearance of uncoerced loyalty and thus has an incentive to hide the repression that disciplines the public sphere. The result is that the authoritarian public

---

<sup>9</sup> For a detailed portrait of the newspaper as an example of the Chinese *Baihua* movement, see Pistorius 2019. Erhard and Hou (2018: 15–17) list twelve Tibetan language newspapers published in the Republic of China but admit that only the *Vernacular News in Tibetan Language* has become accessible.

<sup>10</sup> While the absence of a public sphere in the context of authoritarian or totalitarian states, such as Myanmar or North Korea, is widely accepted, Fiedler and Meyen (2015) nevertheless argue that, e.g., in the GDR, a public sphere, however limited, existed.

sphere is characterized by both state repression and state-manufactured legitimating messages. (Dukalskis 2017: 26)

In Tibetan exile communities, the Central Tibetan Administration (CTA), i.e. the Tibetan government-in-exile based in Dharamsala (H.P) in India, exerted significant control over Tibetan newspapers of the 1960s,<sup>11</sup> as becomes obvious from the recollections of Gönpo Dorje (Mgon po rdo rje, 1935–2016), a former member of the Tibetan parliament-in-exile who had worked since 1964 in various functions for CTA newspapers. He summarises the aims of the first exile newspaper *Freedom* (FRD, *rang dbang gsar shog*):

... recognising the need for communication between the exiled Tibetan government and the Tibetan people, for Tibetans to learn about world news, and for a modern, quality newspaper in the Tibetan language to serve as a platform for Tibetan government politics, initially Gyalo Thondup (Rgya lo don grub, born 1928)<sup>12</sup>, the elder brother of His Holiness the Dalai Lama, took full responsibility. He established the *Freedom Press* in Darjeeling, arranged for funding and materials, purchased the necessary equipment, organised staff, and made all preparations without hesitation. Then, the Kashag (Tibetan cabinet) in Mussoorie appointed Yeshe Dargye, an official of the Tsedrung rank, as the head of the *Freedom Press* in Darjeeling.<sup>13</sup>

---

<sup>11</sup> A free press only developed later from the 1970s on with publications such as the *Tibetan Review* under the editorship of Dawa Norbu, or in the 1990s, the short-lived *Mangtso* edited by Jamyang Norbu, Lhasang Tsering and Tashi Tsering, see, for example, Norbu 2011.

<sup>12</sup> Although Gyalo Thondup did not hold official positions in the CTA at that time, he acted for several decades from the 1950s onwards as an important negotiator between Tibet, the Republic of China (ROC) and the PRC as well as a political agent in the United States, and from the late 1970s (cf. Rgyal lo don grub 2015).

<sup>13</sup> [...] byes 'byor bod gzhung dang / bod mi mang dbar la gnas tshul shes rtogs bya rgyu dang / bod mi rnams kyis 'dzam gling gi gsar gnas shes rtogs bya thabs dang / bod gzhung gi chab srid kyi gleng stegs sogs la bod mi rang nyid kyi skad yig thog nas deng dus dang mthun pa'i gsar shog tshad ldan zhig dgos gal che bar gzigs tel thog mar yab gzhis stag 'tsher gyi sa chen rgya lo don grub mchog nas thugs khur rkang bzhes kyis rdor gling rang dbang gsar khang tshugs yul dang / dngos dngul thabs shes gnang rgyul 'phrul 'khor yo byad spus gzigs gnang rgyul las byed pa go sgrig gnang rgyu sogs sku ngal 'dzem med thog nas gra sgrig cha tshang zin pa dang / ma su ri bzhugs sgar bka' shag nas rtse drung



In the 1960s, Tibetan exiles initially settled in the long-established communities in Darjeeling and Kalimpong in West Bengal, the newly founded settlement in Mussoorie in Uttarakhand, or worked in road construction in Sikkim. Lacking sufficient skills to read news in Hindi or English, it was strongly felt that the scattered community needed “a modern, quality newspaper.” The main protagonists in setting up the newspaper as a means “for communication between the exiled Tibetan government and the Tibetan people” were closely linked to the Dalai Lama and his government.

The objectives of Darjeeling’s *Freedom* were to be a political tool to protect the interests of Tibet as a nation and the religious and political freedoms of the Tibetan people. In the mid-20th century, when Tibetan communities were separated, parents, children, relatives, and spouses were scattered under Chinese communist oppression, not knowing where others had ended up — like blind people lost in the desert. During this time of great hardship for Tibetan refugees, *Freedom* published announcements that helped many families reunite and reconnect. It showed exiled Tibetans how to contact the Tibetan government-in-exile and provided extensive guidance on livelihoods, education, healthcare and other matters. At that critical time when the fate of Tibet as a nation and the Tibetan people hung in the balance, *Freedom* continuously published political guidance for Tibetan communities on how to navigate immediate and long-term challenges, how to distinguish between enemies, friends and protectors, and how to chart a path forward. This benefited both the government and the people.<sup>14</sup>

---

*las tshan ye shes dar rgyas rdor gling rang dbang gsar khang gi 'go 'dzin du bsko dzongs gnang ba* (Mgon po rdo rje 2015: 54).

<sup>14</sup> *Rdor gling rang dbang gsar shog gi dmigs yul ni/ bod rgyal khab dang / bod mi rigs kyi chos srid rang dbang dang bcas pa'i khe phan srung skyob byed thabs chab srid kyi lag cha zhig pa dang / dus rabs nyi shu pa'i dkyil smad tsam la gzhis byes bod mi rnams kha bral gyi dus skabs la/ rgya dmar gyis drag gnon 'og nas pha ma dang / bu phrug gnyen nyel/ bza' tshang sogs kha 'thor nas gar slebs/ gar yod mi shes par long ba thang dkyil du lus pa ltar btson byol bod mi rnams dka' ngal che dus/ rang dbang gsar shog gi thog nas gsal bsgrags byas pa la brten nas/ gnyen ngo 'phrod de khyim tshang mang po rug 'dzoms thub pa dang / byes 'byor bod gzhung la 'brel ba zhu lam bstan pa dang / de bzhin 'tsho thabs dang / shes yon slob grwa/ 'phrod bsten sogs kyi lam ston rgya cher byed thub pa dang / skabs der bod rgyal khab dang / bod mi rigs bcas pa'i gnas stangs 'chi gson gyi bar*

Although the Tibetan societies in the PRC and the Tibetan diaspora are ideologically different – the first being a Leninist party state and the latter moving towards democracy – in the 1950s and 1960s, their use of media in the PRC and in Tibetan communities of the early exile appears to share some similarities.<sup>15</sup> Among the Tibetan exiles, newspapers were a tool to connect the diasporic community, spread out initially across the subcontinent, including Nepal and later globally. Additionally, the newspapers of the CTA were intended to create a Tibetan community that shares a worldview, a set of values, and a specific concept of Tibet, or in other words, create unity. Besides its practical value as a communication channel for the CTA to the Tibetan exile community, the newspapers also served as a community or even Tibetan nation-building tool.<sup>16</sup>

Similarly, Tibetan newspapers in the PRC served as a tool for the Communist government to reach out to its Tibetan subjects and communicate the ruling ideology (mostly unknown and unintelligible for most Tibetans at the time) as well as government incentives, policies and campaigns. At the same time, the newspapers were seen as a tool of persuasion to better integrate Tibetans into the multi-ethnic Chinese society (*krung hwa/zhong hua*).

Although, as mentioned earlier, Tibetan newspapers existed since 1904, with the founding of the PRC and the Chinese annexation of Tibet in the following decades, newspapers became, together with radio broadcasts, the sharpest instruments of propaganda in the China-Tibet conflict. Franklin Houn, writing in the 1960s, summarised the function of the press in the early PRC:

---

*mtshams su slebs brten/ dza drag gi dus skabs der rang dbang gsar shog thog gzhis byes bod mi rnams la 'phral phugs kyi mdun lam dgra gnyen mgon gsum 'dzin stangs ma 'dzol bar byed sgo'i chab srid kyi lam ston rgyun mthud nas gsar spel byas pas gzhung dmangs gnyis phan byung* (Mgon po rdo rje 2015: 55–56).

<sup>15</sup> Roemer (2008: 165–166) discusses the problems faced by Tibetan exiles in their process of democratic transformation and identifies a lack of freedom of speech until the 1990s.

<sup>16</sup> According to McLagan, before the 1990s the CTA censored publications, “for the sake of maintaining unity in exile” (McLagan 1996: 238).

The Chinese government, like that of the Soviet Union, maintains strict controls over the entire publications industry. The press, by its nature peculiarly responsive to changing events, appears to serve the following four functions: propaganda, agitation, public information, and “self-criticism.” (Houn 1961: 91)

Even though the statist and propagandistic nature of Tibetan language newspapers within Tibet or the PRC makes them a problematic source for studying Tibetan society and sentiment, they contain in a relatively unfiltered form the state’s intentions, policies, and concepts, or at least those that the state and the Party considered favourable to its project. They also reveal the state’s strategies of communication to its Tibetan subjects. In this respect, newspapers – and other public documents, such as policy publications and official autobiographies – can be used to make up for the lack of available or accessible archival documents.<sup>17</sup> The DD Corpus of early Tibetan newspapers of the 1950s and 1960s was compiled to provide a resource for those studying this crucial period in Tibetan history.

### *1.1 Tibetan Newspapers in the People’s Republic of China between 1950 and 1965*

In his 1961 study of Chinese Communist propaganda, Franklin Houn describes the Chinese press in the early Maoist period (Houn 1961: 91–154). He describes the Chinese newspaper landscape in horizontal and vertical dimensions. The vertical organisation of the newspapers corresponds with four of the five geographic and administrative levels into which the PRC is divided – national, provincial, prefectural (district), and county – to ensure a newspaper network on each level, “to serve the party, government, and mass organization” (Houn 1961: 103).

---

<sup>17</sup> While archives in the PRC are sometimes accessible, in Tibetan areas research is extremely difficult and access to archives mostly impossible. With regard to exile communities and the CTA, archives are also limited.

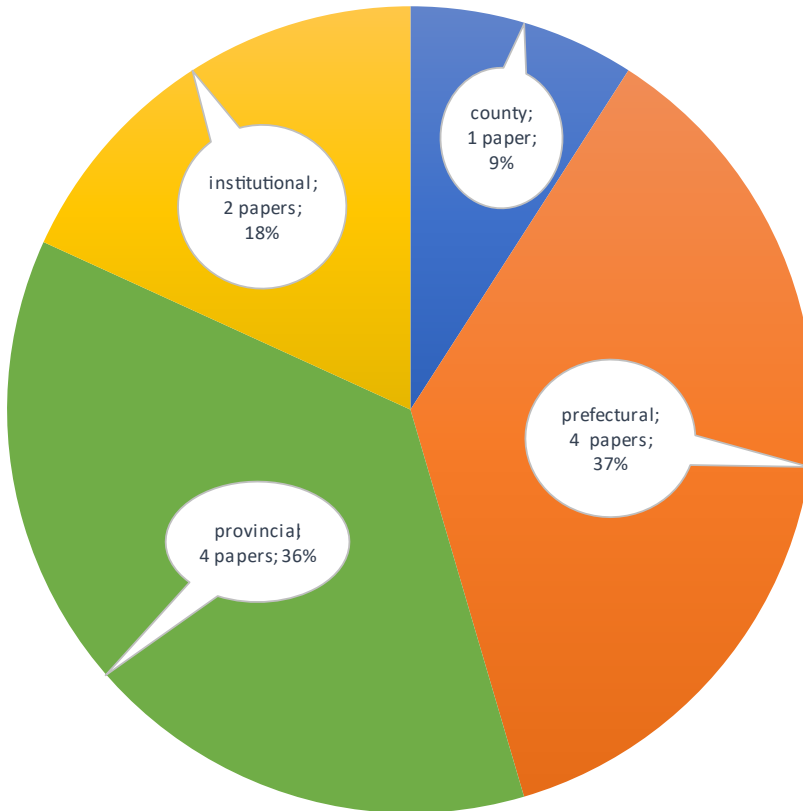


Figure 1 Horizontal division of newspapers corresponding to PRC administrative levels in the *Divergent Discourses* corpus

The horizontal organisation of the newspaper publications follows the principle of specialisation. On each (vertical) level, the state apparatus produces highly specialised publications targeting specific groups, e.g. the youth, university students, workers and other population groups. For the Chinese language press in 1956, Houn explains that 352 newspapers existed above the prefectural level, and 33% were targeting the youth. Of the 1,049 county-level publications, 49% were published for farmers and workers, while special papers were published for industrial and mining centres and targeting workers (Houn 1961: 106–107).

The minority language newspapers follow this pattern of vertical divisions. The DD Corpus includes two provincial-level publications, the *Qinghai Tibetan News* (QTN, *Mtsho sngon bod yig gsar 'gyur*)<sup>18</sup> and the *Tibet Daily* (TID, *Bod ljong nyin re'i gsar 'gyur*). It contains several prefectural-level papers – the *Minjiang News* (MJN, *Ming kyang tshags dpar*), the *South Gansu News* (SGN, *Kan lho gsar 'gyur*), and the *Ganze Daily News* (GDN, *Dkar mdzes nyin re'i gsar 'gyur*) – and one county-level newspaper, the *Gyantse Daily News* (GTN, *Rgyal rtse nyin re'i gsar 'gyur*).

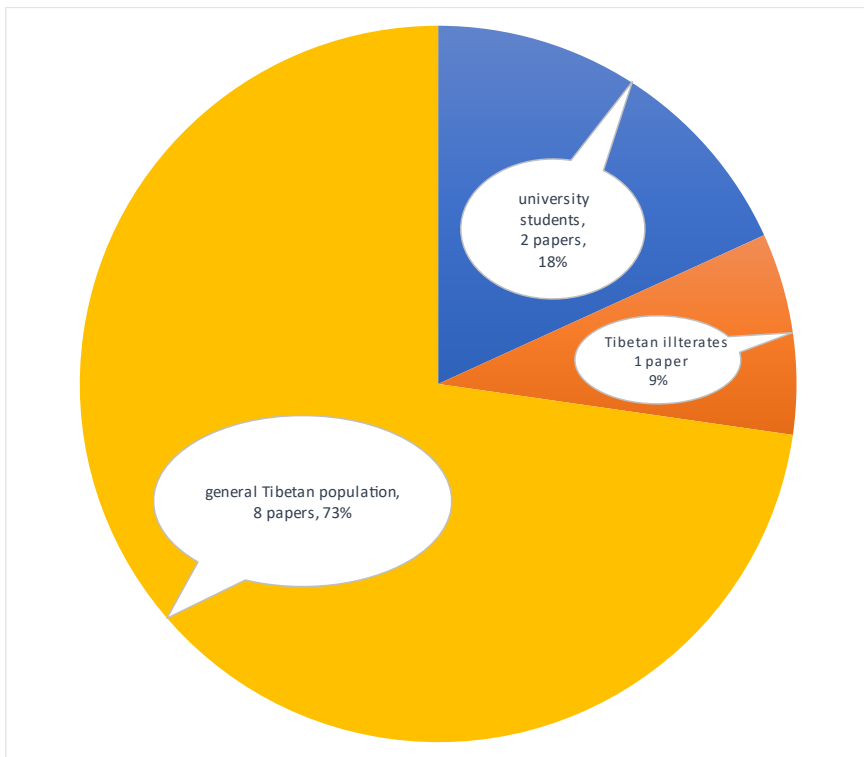


Figure 2 Functional division of Tibetan newspapers in the PRC in the Divergent Discourses corpus

<sup>18</sup> The three-letter sigla for each newspaper title are used in the filenames for each image of a newspaper page in the Diverge corpus to facilitate identifying the newspaper title.

Most newspapers were published for the general Tibetan population. However, two newspapers in the corpus do not correspond with an administrative region. These are university newspapers published for students at one or another of the “Institutes of Nationalities” (*mi rigs slob chen*; Chin. *minzu xueyuan*). The *Tibet Daily Pictorial* (TDP; *Bod ljong nyin re'i gsar 'gyur brnyan par*) is the only example in the corpus of publications that targets a less educated and illiterate population by focussing on visual communication in the form of photographic reproductions, drawings, or cartoons.

While Tibetan newspaper publications serve at the same time as functional publications, ensuring smooth and (in theory) culturally sensitive communication with the Tibetan population, the Divergent Discourses corpus contains no examples of functional publications targeting a specific group, such as workers, soldiers, students, or farmers, among the Tibetan population, though these were commonly published in the reform period after 1978.

It seems likely that the creation of a Tibetan-language newspaper network on a scale or form similar to that of Chinese-language newspapers was challenged by several factors, including the economy and demography of Tibetan areas.

### 1.2 *Tibetan Newspapers published in exile communities in India*

Being less organised and lacking strong state central control, Tibetan newspapers published in India appear to be more diverse than their trans-Himalayan counterparts. Although smaller in number, we have a more complete picture of their publication history. All the newspapers we identified targeted the scattered Tibetan community in India and beyond. As such, these newspapers would correspond to national-level newspapers like the Chinese-language *People's Daily* (*Renmin ribao*) in the PRC.

Unlike the fully state-controlled Communist hierarchical structure found in the PRC, we find in India a liberal model in the early years that gradually gives way over time to a more centralised one, controlling the relationship between the administration and the press.

Until the arrival of the Dalai Lama and the first wave of Tibetans as refugees in India, besides the very local papers published by Moravian missionaries, the only Tibetan newspaper was Dorje Tharchin's (1890–1976) privately published the *Tibet Mirror* (TIM, *yul phyogs so so'i gsar 'gyur me long*), which ran from 1925 to 1963. As an inde-

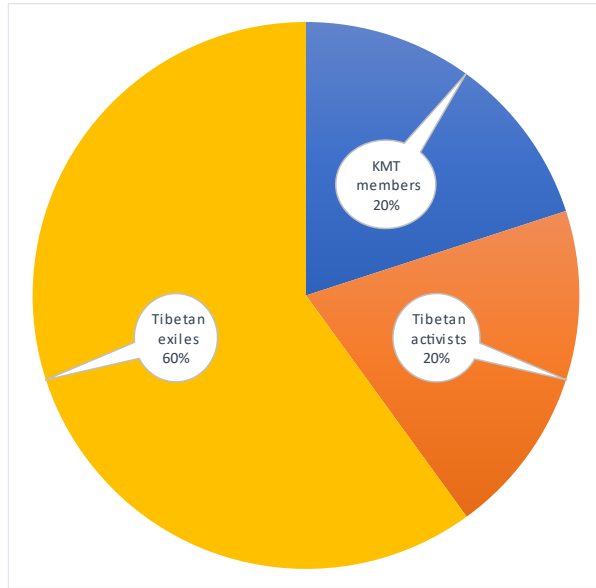


Figure 3 Functional division of Tibetan newspapers in India in the Divergent Dis-courses Corpus and their target audience.

pendently run newspaper, the *Tibet Mirror* was undoubtedly influenced by various actors, such as the British Indian government and later the Indian government. Still, editorial decisions were always taken by the editor, Tharchin (Moskaleva 2023: 47–55).

With the gradual establishment of Tibetan refugee settlements and, later, of the Tibetan exile administration, the need for effective communication was felt, and Tibetan newspapers were founded in the West Bengal hill station of Darjeeling. The CTA could not exercise media power or control on the scale of China, but it still appointed the editorial office and funded the publication. The first paper, *Freedom*, was started in March 1960, just one year after the 1959 uprising. It was followed by a short-lived second newspaper, *Defend Tibet's Freedom* (DTF, *rang dbang srung skyob gsar shog*), run by Lhamo Tsering (1924–1999), who later became Gyalo Thondup's deputy in organising the guerrilla forces. In 1965, both papers were ordered by the CTA to merge. They then formed the daily newspaper *Tibetan Freedom* (TIF, *bod mi'i rang dbang*), which, from 1978 onwards, became more closely integrated into the CTA as an official gazette and is still running today,

edited by the Department of Information and International Relations (DIIR) in Dharamsala. These publications served as the main means of communication for the CTA with the general Tibetan exile public.

At the same time, there were at least two unrelated Tibetan-language newspaper projects in India. One was a party newspaper called *Central Weekly News* (CWN, *krung dbyang gsar 'gyur*). It was published by the Chinese Nationalist Party then based in Taiwan, the Kuomintang (KMT), and targeted a Tibetan exile audience from 1960 onwards. In the 1950s and early 1960s, the KMT as an opponent of the Chinese Communist Party appealed to some Tibetans as a potential ally in their quest for independence (although in fact the KMT was opposed to independence for Tibet).<sup>19</sup>

The *Central Weekly News* was published by Tempa Landoo (Bstan pa lhun grub, 丹巴隆舟 Danba Longzhou), also known by his Chinese name, Gao Qiangui 高攀桂, until the late 1970s in Calcutta.<sup>20</sup>

The other Tibetan-language paper from this period had no title. It was issued under the name of *News Office of the Indian Government (Gangtok)* and published in the mid-1950s. Only a few copies have survived in the Grassi Museum für Völkerkunde (VM) but these are currently misplaced and thus inaccessible.

In sum, the development of Tibetan newspaper publishing in India until 1965 went from a proselytising Moravian press to a long-running independent modern newspaper, *The Tibet Mirror* and eventually to a government gazette, the *Tibetan Freedom*. The rise of the Communist Party in China and its integration of Tibet in the People's Republic in the 1950s resulted in the rise of institution-controlled newspaper

---

<sup>19</sup> Moskaleva describes how, for example “in Tharchin’s anti-communist discourse, the Kuomintang government occupies an important position of a ‘friend’ of Tibet.” If successfully overthrowing Communist rule in China, “the Kuomintang promises to grant independence to Tibet” (2023: 274).

<sup>20</sup> I am grateful to Ling-wei Kung (Taipei) for clarification of Tempa Landoo’s Chinese name and identity. He is still best known as a KMT member and translator of Sun Yat-sen’s (孫中山 1866–1925) *Three Principles of the People Dmangs gsum ring lugs* (三民主義 *sanmin zhuyi*), which he first published in 1974 at the *Krung dbyang gsar 'gyur par khang* in Calcutta. I want to thank Chen Nai-hua for sending me images of the edition Sun 1985.



projects in exile, whether run by actors such as the Indian branch of the KMT, the Indian or Sikkimese government, or the CTA.

## 2 *Acquisition of Materials*

Early Tibetan newspapers until 1965 are widely dispersed globally; no single institution or library holds a comprehensive collection of Tibetan newspapers.<sup>21</sup>

The idea of compiling a larger corpus of such papers and making it available in digital form grew out of the rediscovery of the newspapers brought back by Johannes Schubert (1896-1976) in 1955 and given to Leipzig's Grassi Museum für Völkerkunde (MV)<sup>22</sup> and those collected by Josef Kolmaš (1933-2021) between 1957 and 1959 during his stay in Beijing and given to the Oriental Institute (OI) of the Czech Academy of Sciences in Prague.<sup>23</sup> The two collections combined provide a comprehensive collection of newspapers – although incomplete – from 1954 to 1958.

A digital collection containing ca 70% of the entire *Tibet Mirror* print run was compiled, digitised and made openly available under the guidance of Luran Hartley at Columbia University (CU) from 2009 to 2013.<sup>24</sup> This collection combined the 97 issues of the Columbia University Library's Tharchin Collection with the *Tibet Mirror*

---

<sup>21</sup> For post-Cultural Revolution, reform era newspapers, the situation is different and Tibetan language newspapers are held at Columbia University, Staatsbibliothek zu Berlin, the Library of Congress and others.

<sup>22</sup> Schubert published a descriptive catalogue of his acquisitions in 1958. Most publications still uncatalogued are in the library of the Grassi Museum für Völkerkunde in Leipzig. The sigla in parenthesis are used in the filenames of the individual newspaper pages to allow the persistent identification of the library where the newspapers are held.

<sup>23</sup> Kolmaš published a descriptive catalogue of this collection in 1978.

<sup>24</sup> [https://archive.org/details/ldpd\\_6981643\\_000](https://archive.org/details/ldpd_6981643_000) (accessed on January 10, 2025); for a detailed description of the Tharchin Collection, including the digital holdings of *Tibet Mirror* see [https://library.columbia.edu/libraries/eastasian/special\\_collections/tibetan-rare-books---special-collections/tharchin.html](https://library.columbia.edu/libraries/eastasian/special_collections/tibetan-rare-books---special-collections/tharchin.html) (accessed on January 10, 2025).

holdings of Yale University's Beinecke Rare Book and Manuscript Library<sup>25</sup> in the United States, and the collections of the Collège de France (CF)<sup>26</sup> and the Musée Guimet.<sup>27</sup>

Another effort to digitise and transcribe the *Tibet Mirror* was undertaken from 2015 to 2018 by the Collège de France (Wang-Toutain et al. 2018: vii; Wang-Toutain 2018). The project was joined by the Library of Tibetan Works and Archives (LT), which holds another collection of 254 issues of the *Tibet Mirror* (Topgyal 2016).

As noted by Anna Sawerthal in her work on the *Tibet Mirror* (Sawerthal 2018: 343–346), the library of the Department for Indology, Tibet and Buddhism (IT) at Vienna University holds a significant collection of the main five Tibetan-language newspapers published in the early years of Tibetan exile in India, including the *Tibet Mirror*, *Freedom*, *Defend Tibet's Freedom*, *Tibetan Freedom* (TIF, *bod mi'i rang dbang*), and *Central Weekly News*. These collections are complemented by the holdings of *Defend Tibet's Freedom* at the British Library (BL) in London, of *Freedom* in the Bodleian Library (BD), Oxford, and those of the *Central Weekly News*, held at the library of the University of Washington (UW) and the Library of National Chengchi University (NC), Taipei. There are also copies of some of these titles in the holdings of the Library of Tibetan Works and Archives, the Tibet Museum (TM), and a few copies are also in private collections. Other papers were already fading from collective memory. The historically interesting *Defend Tibet's Freedom*, which was started by Gyalo Dhondup in 1963 and run by his close assistant Lhamo Tsering with funding from the CIA, is virtually unknown, and seemingly no copy has survived in the exile community in India. At least one other Tibetan-language newspaper is known to have existed at this period: the publication produced for Tibetan guerrilla forces based secretly in

<sup>25</sup> <http://beinecke.library.yale.edu/digitallibrary/tibetmirror.html> and <https://collections.library.yale.edu/catalog/2057570> (accessed on January 10, 2025).

<sup>26</sup> [https://omnia.college-de-france.fr/permalink/33CDF\\_INST/1kslc0r/alma990004692210107166](https://omnia.college-de-france.fr/permalink/33CDF_INST/1kslc0r/alma990004692210107166) and [https://salamandre.college-de-france.fr/archives-en-ligne/ead.html?id=FR075CDF\\_00IET00TM&c=FR075CDF\\_00IET00TM\\_e0000002&qid=eas1736260177978](https://salamandre.college-de-france.fr/archives-en-ligne/ead.html?id=FR075CDF_00IET00TM&c=FR075CDF_00IET00TM_e0000002&qid=eas1736260177978) (accessed on January 10, 2025).

<sup>27</sup> <http://www.guimet.fr/fr/> (accessed on January 10, 2025).

Mustang, Nepal, until their dissolution in the mid-1970s. Called *Understanding* (GOT, *go rtogs*), it seems that no copies have ever made it into the archives.

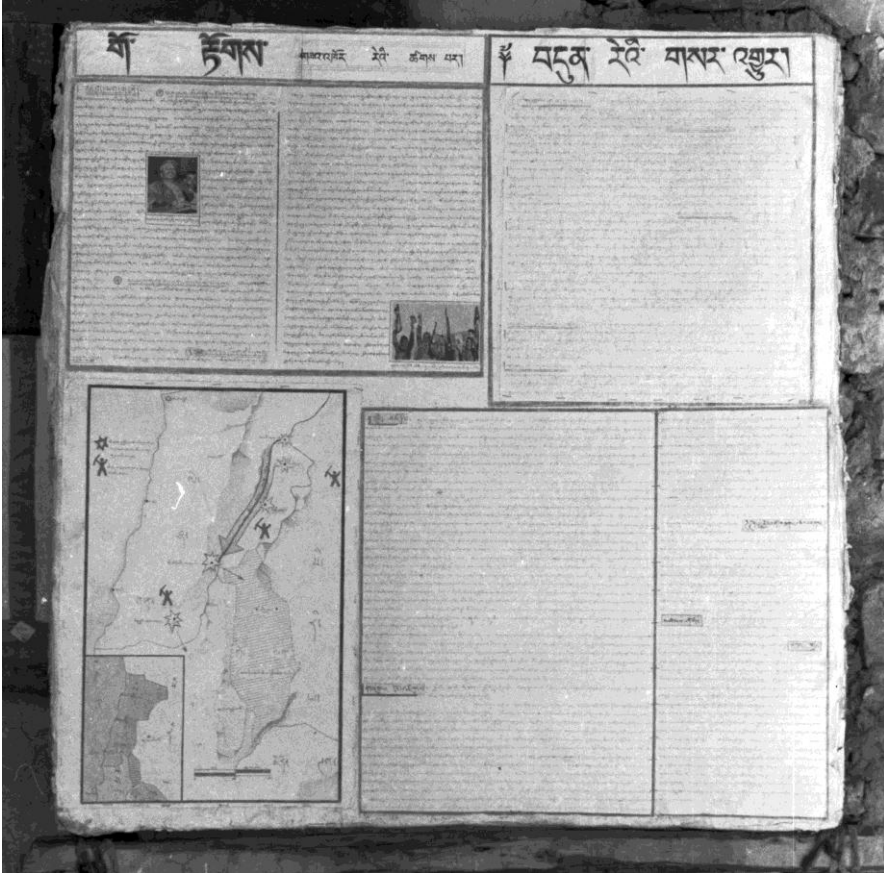


Figure 4 Issue of *Understanding* (*go rtogs*) put up on a wall (Lhamo Tsering Archive / White Crane Films)

As for papers from within Tibet, several libraries in Europe and the US have microfilm editions of the *Qinghai Tibetan News* and of the *Tibet Daily*.<sup>28</sup> Smaller collections of original copies of these provincial-level newspapers from within China are in the Schubert and Kolmaš

<sup>28</sup> The microfilm editions held by the Columbia University Library (CU), the Library of Congress, appear to be the same as the copies held at the Staatsbibliothek zu Berlin (SB) produced by China National Microforms Import & Export Corporation, Beijing.

collections in Leipzig and Prague, respectively, and in the libraries of Columbia University and Vienna University.

This survey of available materials outside China demonstrates how widely scattered the remaining copies of Tibetan newspapers from that period are. In addition, there are several newspaper titles whose existence is known primarily from hearsay, mainly smaller newspapers published in outlying districts of Tibet. For instance, our corpus has a single issue of such a newspaper, the *Gyantse News*. We can assume, there was also a *Shigatse Newspaper* and perhaps a *Nagchu Newspaper*, but no issues of such papers from the early years have come to light.<sup>29</sup>

Besides the corpus's academic value, we hope that compiling digital copies of these widely dispersed papers will enable this archive to be preserved and accessible to the wider Tibetan community.

### 3 *The Divergent Discourses Corpus of Tibetan Newspapers from 1950 to 1965*

The DD Corpus combines in one collection eleven newspapers published within the PRC and five newspapers published in exile communities on the Indian subcontinent.<sup>30</sup>

---

<sup>29</sup> In fact, in the 2000s and 2010s, among others, both a *Nagchu Newspaper* (Nag chu tshags par 那曲报 *Naqu bao*) and *Shigatse Newspaper* (Gzhis rtse'i tshags par 日喀则报 *Rikaze bao*) were published (Erhard 2015: 166–167). See also Hartley 2005 for an overview of newspapers in the post-Mao era.

<sup>30</sup> Although we know of them, newspapers of the 1950s such as the རྒྱ་གར་གཞུང་གི་གསར་འཕེལ་གྱི་ལུང་ལྷན་ཁང་གི་སྐད་ལྗང་ (rgya gar gzhung gi gsar spel las khungs sgang thog) published in Gangtok, Sikkim, or the ཀུན་མཁྱེན་རྒྱལ་ཁྲིམས་ལྷན་ཁང་གི་སྐད་ལྗང་ (kun khyab rlung 'phrin gsar 'gyur) published in Beijing, we were not able to obtain copies so far and hence could not include them in the corpus.

## 3.1 Newspaper Publications from India in the DD Corpus

	Newspaper title	Place of publication	Years in DD Corpus	Issues/ pages in DD Corpus
1.	ཀླུང་དབྱུང་གསལ་འགྲུལ ( <i>krung dbyang gsar 'gyur</i> ) 中央週報 ( <i>zhong yang zhou bao</i> ) "Central Weekly News" (CWN)	Calcutta, W.-Bengal	1963–1964	30 issues 120 pages
2.	རང་དབང་སྲུང་སྐྱོབ་གསལ་ཤོག ( <i>rang dbang srung skyob gsar shog</i> ) "Defend Tibet's Freedom" (DTF)	Darjeeling, W.-Bengal	1963	14 issues 240 pages
3.	རང་དབང་གསལ་ཤོག ( <i>rang dbang gsar shog</i> ) "Freedom" (FRD)	Darjeeling, W.-Bengal	1961–1965	203 issues 1420 pages
4.	བོད་མིའི་རང་དབང་ ( <i>bod mi'i rang dbang</i> ) "Tibetan Freedom" (TIF)	Darjeeling, W.-Bengal	1965	294 issues 916 pages
5.	ཡུལ་ཕྱོགས་སོ་སོའི་གསལ་འགྲུལ་མེ་ལོང་ ( <i>yul phyogs so so'i gsar 'gyur me long</i> ) "Tibet Mirror" (TIM)	Kālimpong, W.-Bengal	1950 – 1963	97 issues 1008 pages

## 3.2 Newspaper Publications from the PRC in the DD Corpus

	Newspaper title	Place of publication	Years in DD Corpus	Issues/ pages in DD Corpus
6.	དཀར་མཛེས་ཉིན་རེའི་གསར་འགྲུར ( <i>dkar mdzes nyin re'i gsar 'gyur</i> ) 甘孜日報 ( <i>ganzi ribao</i> ) Ganze Daily (GDN)	Dartsedo, Sichuan	1959	74 issues 296 pages
7.	རྒྱལ་ཤེགས་འགྲུར ( <i>rgyal rtse gsar 'gyur</i> ) 江孜報 ( <i>jiangzi bao</i> ) Gyangtse Daily News (GTN)	Gyangtse, TAR	1954	1 issue 4 pages
8.	དར་མདོའི་གསར་འགྲུར ( <i>dar mdo'i gsar 'gyur</i> ) 康定報 ( <i>kangding bao</i> ) Kangding News (KDN)	Dartsedo, Xikang <sup>31</sup>	1954–1955	15 issues 60 pages
9.	མིང་ཀླང་ཚགས་དཔར ( <i>ming kyāng tshags dpar</i> ) 岷江報 ( <i>minjiang bao</i> ) Minjiang River News (MJN)	Maoxian, Sichuan	1953–1955, 1959	83 issues 348 pages
10.	གསར་འགྲུར་མདོར་བསྡུས ( <i>gsar 'gyur mdor bsdus</i> ) 新聞簡訊 ( <i>xinwen jianxun</i> ) News in Brief (NIB) <sup>32</sup>	Lhasa, TAR	1953–1955	96 issues 394 pages

<sup>31</sup> The province Xikang 西康, or Shis khams in Tibetan, was established in 1939 and dissolved into the Tibetan Autonomous Region (TAR) and Sichuan in 1955.

<sup>32</sup> Bapa Phüntso Wangye in his (auto)biography mentions “Brief Communications in Tibetan” or *bod yig bsdus 'phrin*, an early 1950s newsheet edited in 1952 by “a new research committee that translated news and directives into Tibetan”, as the predecessor of the Tibet Daily (Goldstein 2004: 179). Its title and description are similar to NIB (= *Gsar 'gyur mdor bsdus*), but predate it by at least one year. It is

	Newspaper title	Place of publication	Years in DD Corpus	Issues/pages in DD Corpus
11.	མཚོ་སྔོན་བོད་ཡིག་གསར་འགྲུར ( <i>mtsho sngon bod yig gsar 'gyur</i> ) 青海藏文報 ( <i>qinghai zangwen bao</i> ) Qinghai Tibetan News (QTN)	Xining, Qinghai	1951-1960, 1963-1965	1,083 issues 4,098 pp.
12.	ཀན་ལྷོ་གསར་འགྲུར ( <i>kan lho gsar 'gyur</i> ) 甘南報 ( <i>gannan bao</i> ) South Gansu News (SGN)	Zö (Tib. gtsos, Chin. Hezuo), Gansu	1959	17 issues 68 pages
13.	བོད་ལྗོངས་ཉིན་རེའི་གསར་འགྲུར་པར་རིས་ ( <i>bod ljongs nyin re'i gsar 'gyur par ris</i> ) 西藏日報 ( <i>xizang ribao</i> ) Tibet Daily Pictorial (TDP) <sup>33</sup>	Lhasa, TAR	1959	2 issues 8 pages
14.	བོད་ལྗོངས་ཉིན་རེའི་གསར་འགྲུར ( <i>bod ljongs nyin re'i gsar 'gyur</i> ) 西藏日報 ( <i>xizang ribao</i> ) Tibet Daily (TID)	Lhasa, TAR	1958-1959, 1961-1965	1980 issues 7641 pages
15.	ལྷོ་རུབ་མི་རིགས་སློབ་གྲྭ་ཆེན་མོ་ ( <i>lho nub mi rigs slob grwa chen mo</i> ) 西南民族學院 ( <i>xinan minzu xueyuan</i> ) South-West Institute for Nationalities (XMX)	Chengdu, Sichuan	1955, 1959	13 issues 52 pages

conceivable that the *Bod yig bsdus 'phrin* represents an earlier name of the *Gsar 'gyur mdor bsdus*.

<sup>33</sup> This newspaper publication is similar yet different from the magazine *China Pictorial* (人民畫報 *renmin huabao*) published in Beijing in various languages, including Tibetan (*mi dmangs brnyan par*) until today. Kamil Sedláček (1972) used *China Pictorial* in his *Tibetan Newspaper Reader*.

	Newspaper title	Place of publication	Years in DD Corpus	Issues/ pages in DD Corpus
16.	<p>ཀླུང་དབྱུང་མི་རིགས་སློབ་གྲྭ།</p> <p>(<i>krung dbyang mi rigs slob grwa</i>)</p> <p>中央民族學院 (<i>zhongyang minzu xueyuan</i>)</p> <p>Central Institute for Nationalities (ZMX)</p>	Beijing	1959	3 issues 10 pages

#### 4 Representation and Bias

The question of representativeness and bias in the form of political representation is at the heart of the Divergent Discourses project, as it is broadly assumed that publications stemming from the PRC will show a Communist, pro-Chinese leaning. In contrast, publications from the subcontinent will have a nationalist and pro-independence perspective. Given that newspapers in the PRC are statist papers, it is important to emphasise that these papers present, first of all, the policies, opinions, and objectives of the state and its administrative apparatus.

It is difficult to conclude who might have been the readership of these publications, but, during high communism in the Maoist period, reading newspapers (or listening in groups to recitations from them) was compulsory communal practice, and the reading public depended on newspapers to indicate

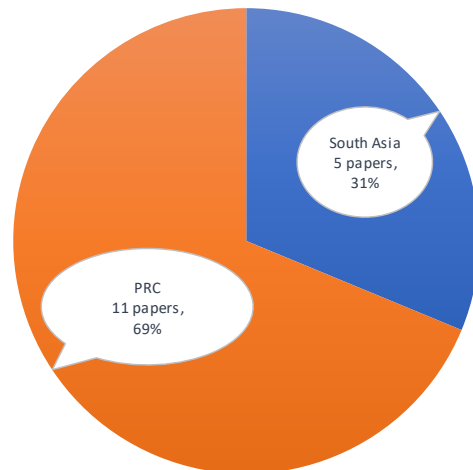


Figure 5 Distribution of PRC (red) and South Asian (blue) newspapers in the DD Corpus



political developments. Nevertheless, these newspapers can be used as sources to study the official narratives and state policies in unfiltered form as they were first communicated to the Tibetan people and evolved over time.

Leaving these fundamental biases aside, collection practices and technical decisions introduce additional bias into the corpus. For example, low quality paper or ink in the original document can result in inferior preservation and can render the text illegible. In such cases, the legibility of the digital version could be of extremely low quality. This was a significant problem with the pages we scanned from a widely circulated microfilm edition. It became apparent that the microfilm was of inferior quality; many pages were photographed twice, some pages were absent, and in many cases, large sections of a page were overexposed. These problems arose with the two daily papers from inside China, the *Tibet Daily* and the *Qinghai Tibetan News*, and image enhancement software was used wherever possible to recover obscured text sections.

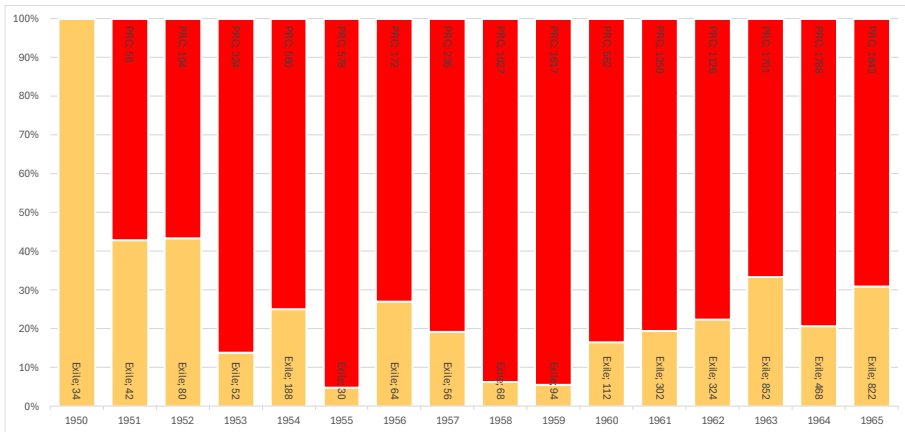


Figure 6 Representation of newspapers in pages from the PRC (red) and from exile (blue) in the DD Corpus over time

Studies of British digital newspaper collections have found that even paper quality – and hence subsequent legibility – can introduce bias into an archive: conservative papers in the UK in the 18<sup>th</sup> and 19<sup>th</sup> centuries, for example, were printed on better paper than left-leaning publications, and so are overrepresented in the archive. Such issues, whether of reproduction quality or of gaps in the archive, are present

with every digital collection, as Beelen *et al.* (2023) have noted, and are important reminders of the need to acknowledge bias and to question whose voices are represented in the pages of these publications.

#### 4.1 Synchronic Aspects: Regional Bias

Among the Tibetan language newspapers published in the early PRC, we recognise a hierarchical stratification of the publications following top-down administrative levels: (1) regional (province/autonomous region), (2) prefectural and (3) county newspapers as outlined earlier. On the provincial level, publications for Qinghai and Tibet/TAR are

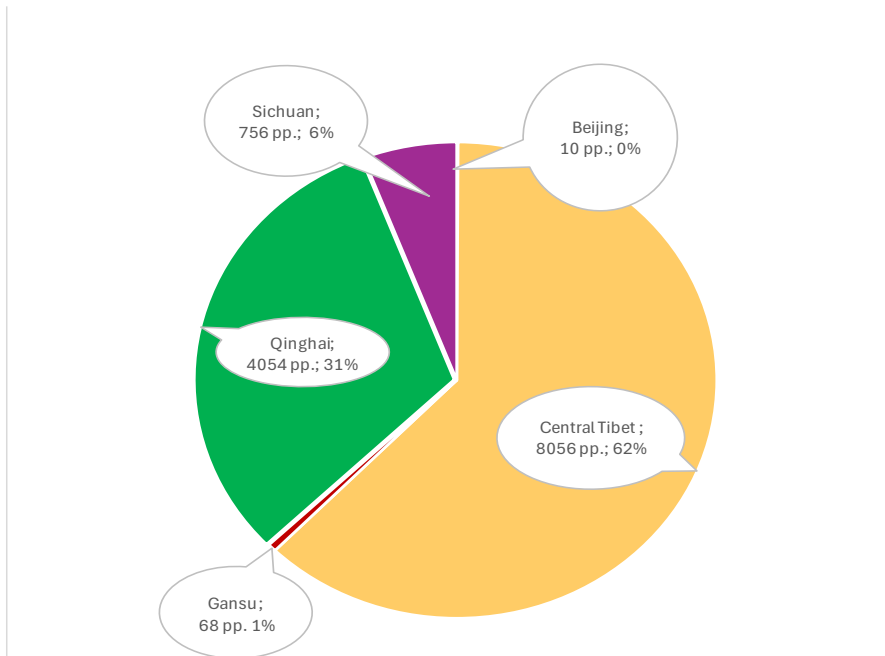


Figure 7 Newspapers in the DD Corpus (in pages) according to PRC province or region of publication

included, while the prefectural-level newspapers all come from Gansu and Sichuan<sup>34</sup>. On the county level, we could only collect a single

---

To keep this simple, the numbers for the former province of Xikang are included in the numbers for Sichuan. For one, there are only very few papers in the corpus

*Gyantse Daily News* issue. The DD Corpus thus reflects the hierarchical stratification of newspaper publications in the PRC. It, nevertheless, is not representative of the newspapers published on each administrative level. County-level publications are in a stark minority while they would have been expected to comprise the largest share of newspaper publications. Similarly, the only prefectural-level newspapers in the corpus are from Ganan (SGN) and Sichuan (MJN,

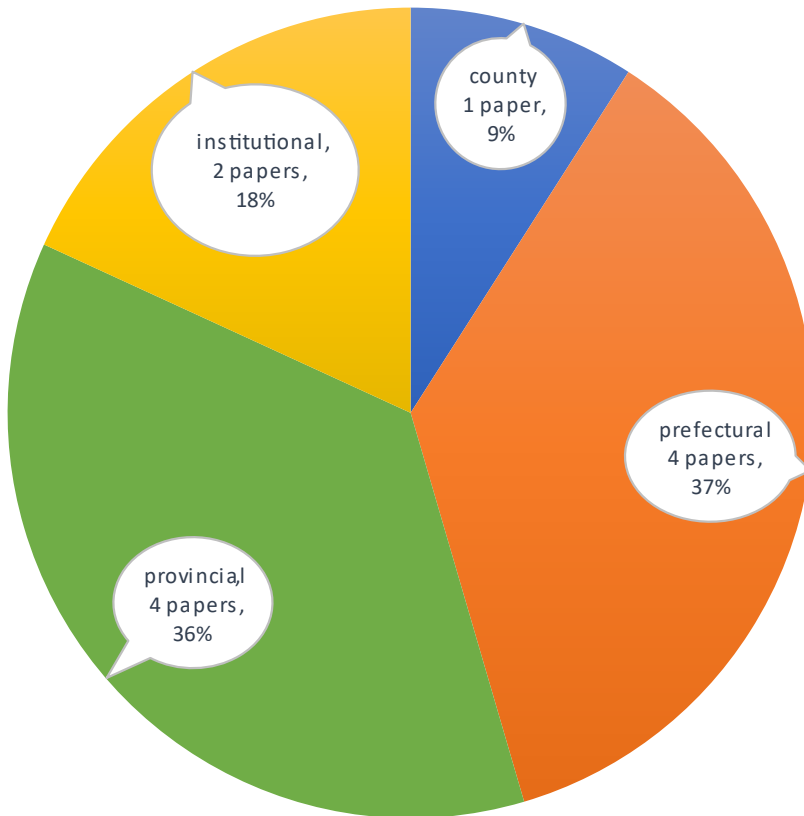


Figure 8 Hierarchical division of newspapers in the DD Corpus in the PRC

published in Xikang. Second, these were then merged or continued as publications from Sichuan, while we do not have any publications in the Diverge Corpus that would have been published in parts of Xikang later coming under TAR administration.

GDN, KDN), while prefectural-level newspapers from Tibet or Qinghai either have not existed at the time or have not been preserved.

Breaking up the newspapers according to the administrative provinces/regions of the PRC makes an even stronger representational imbalance visible. With 68%, most of the corpus (measured in newspaper pages) was published in Tibet/the TAR, and 32% came from Qinghai, while a mere 6% was published in Sichuan and Gansu. The Central Institute for Nationalities (Zhongyang Minzu Xueyuan) in Beijing published a newspaper of which three issues with a total of 10 pages are included in the corpus.

This reveals a significant bias towards publications from Central Tibet and the Tibetan Autonomous Region. The corpus contains three newspapers from Tibet, including the *Tibet Daily*, which, as a daily newspaper, contributed significantly more pages to the corpus than, e.g. the *Qinghai Tibetan News*, which was only published at first weekly and later twice a week.

The regional biases in the Divergent Discourses corpus are important as there are significant differences in policies and their implementations between provinces. Between January and March 1959, for example, the QTN extensively reported on the successful implementation of various policies to technologically advance agriculture and animal husbandry. The majority of articles focus on manure collection, an increase in agricultural production, and collectivisation, while the newspapers published in Maoxian, Sichuan, and Kanze engage in heated class struggles against counter-revolutionaries, monks and former landowners. At the same time, in Tibet before 1959, an altogether different set of policies was implemented. The DD Corpus's stronger representation of Central Tibetan newspapers results in a bias towards more gradual, liberal and less radical attitudes and policies in Tibet.

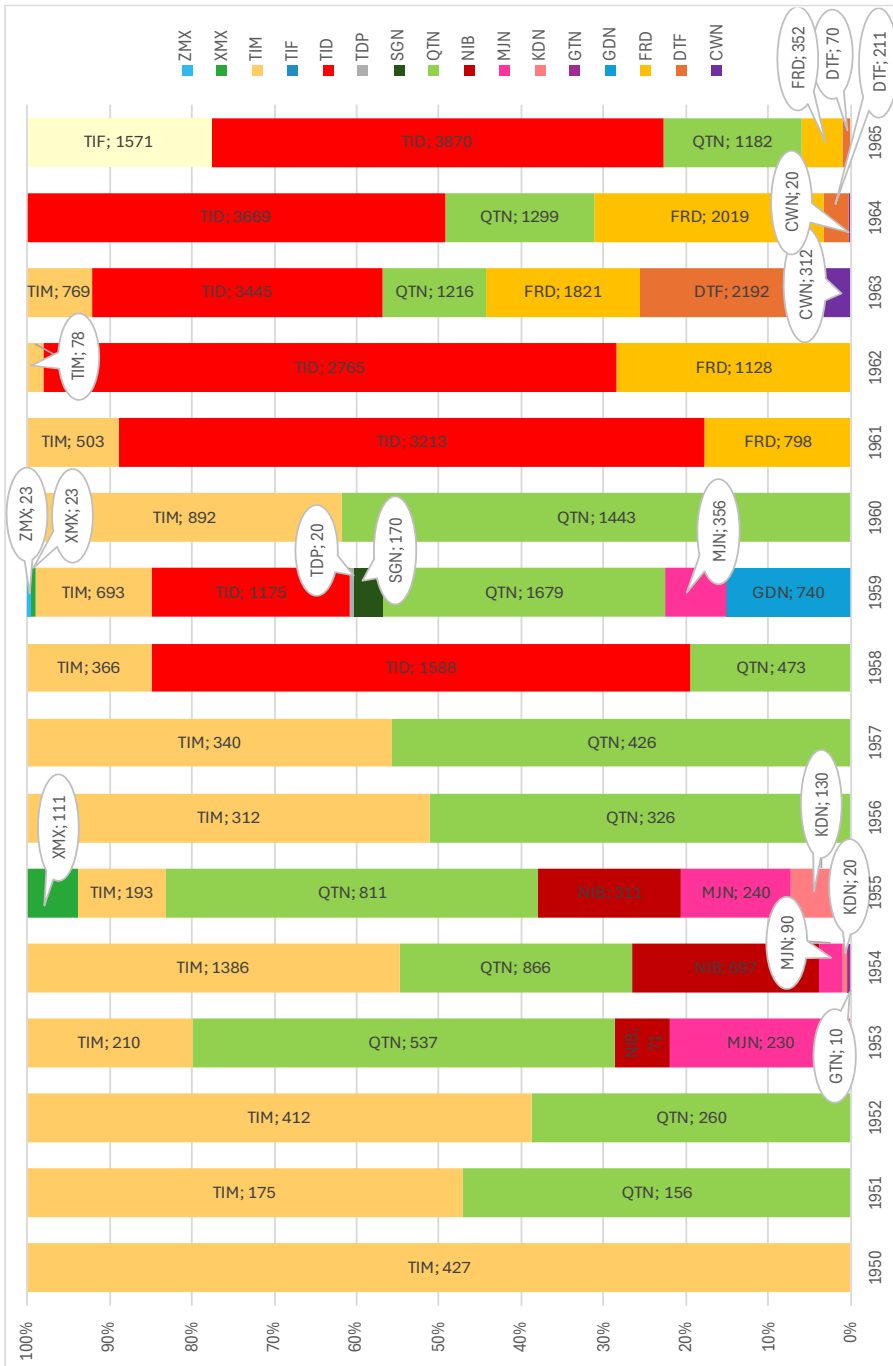


Figure 9 Newspaper shares in pages in the DD Corpus per year.

Newspapers were a relatively new phenomenon in Tibetan areas, and their technological and economic prerequisites were limited. For example, the first Tibetan language newspaper published in the PRC was the QTN in 1951. This was possible because it used a printing press sacked from the Kuomintang in Xining (Hartley 2003: 83). For the year 1950, the Diverge Corpus contains only the *Tibet Mirror*. The following year, with the founding of the QTN, there are two newspapers in the corpus. In 1953, more newspapers were started; the *News in Brief* in Lhasa and the *Minjiang News* in Sichuan. For 1954 and 1955, the DD Corpus features six newspapers from all Tibetan regions, including TIM from Kalimpong in West Bengal, India.

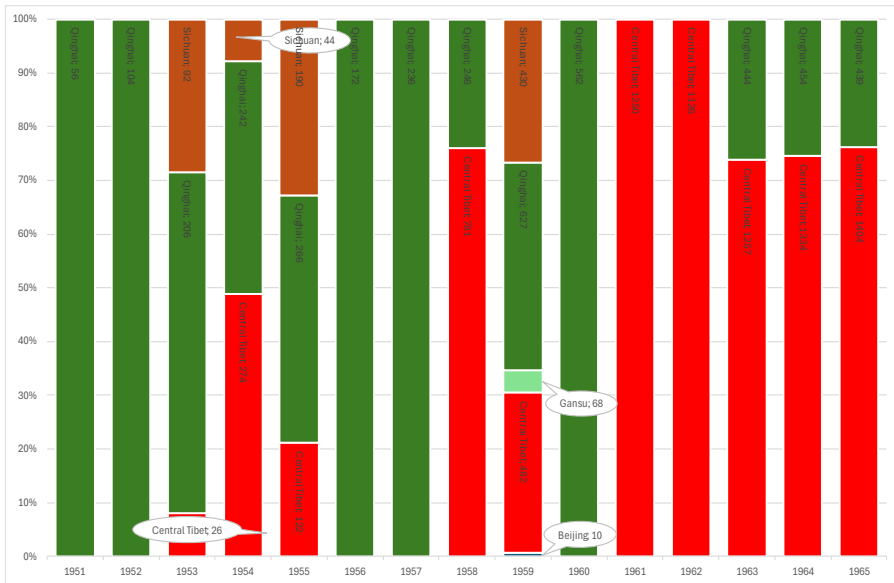


Figure 10 Newspapers in pages per province/autonomous region (PRC) in the DD Corpus per year.

The voice of the exiled Tibetan elite has only been present since 1960 when *Freedom* started publication in Darjeeling. However, from the same year onwards, the representation of PRC newspapers becomes less diverse, mainly the TID from Lhasa and the QTN from Xining.

From Fig 9, it is evident that the Diverge Corpus has significant regional and temporal gaps. Nevertheless, carefully selecting newspapers in the corpus will allow researchers to conduct comparative studies over the whole period. Combining TIM with FRD

and *Tibetan Freedom* allows the study of exile Tibetan positions for the whole 15 years covered in the corpus. Fig. 10 shows that combining *News in Brief* with *Tibet Daily* gives a relatively complete collection from Central Tibet lacking issues for only five years (1950-1952, 1956-1957) which can be compared to the *Qinghai Tibetan News*, which is missing only three years (1950, 1961, 1962) for the whole period, for a diachronic comparative study of Tibet/TAR and Qinghai province.

## 5 Conclusion

The Divergent Discourse Corpus of Tibetan language newspapers is a first attempt to provide a Tibetan language newspaper corpus as a comprehensive historical source. The problematic accessibility of archives and the preservation status and sometimes inferior quality of the original material are reflected in the composition of the corpus. The research design must leverage the resulting representational issues and biases to get representative findings.

Despite the described limitations, the scope of the Diverge Corpus will allow researchers to systematically study the narratives that framed the diverging conceptualisations of Tibet at the time. It will allow – perhaps for the first time – to inquire into the linguistic, social, political and ideological transformations of the 1950s and 1960s.

## Bibliography

- Beelen, Kaspar; Jon Lawrence, Daniel C. S. Wilson, and David Beavan  
“Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan,” *Digital Scholarship in the Humanities* 38(1): 2023, pp. 1–22. [doi:10.1093/lc/fqac037](https://doi.org/10.1093/lc/fqac037).
- Dukalskis, Alexander  
*The Authoritarian Public Sphere: Legitimation and Autocratic Power in North Korea, Burma, and China*. Routledge studies on comparative

Asian politics. London: Routledge, 2017. [doi:10.4324/9781315455532](https://doi.org/10.4324/9781315455532).

Erhard, Franz Xaver

"Tibetan Mass Media. A Preliminary Survey of Tibetan Language Newspapers." In Olaf Czaja and Guntram Hazod (eds.) *The Illuminating Mirror: Tibetan Studies in Honour of Per K. Sørensen on the Occasion of his 65<sup>th</sup> Birthday*, 155–171. Wiesbaden: Reichert, 2015.

"August Hermann Francke (1870-1930), die *Ladakh Agbar* und die ersten tibetischen Zeitungen," *Unitas Fratrum* 80, 2021, pp. 269–90.

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*, 1–40. Paris: Collège de France, 2018.

Fiedler, Anke and Michael Meyen

"'The Totalitarian Destruction of the Public Sphere?' Newspapers and Structures of Public Communication in Socialist Countries: the Example of the German Democratic Republic." *Media, Culture & Society* 37 (6), 2015, pp. 834–49.

Goldstein, Melvyn C.

*A Tibetan Revolutionary: The Political Life and Times of Bapa Phüntso Wangye*. Berkeley, CA: University of California Press, 2004.

Habermas, Jürgen

"The Public Sphere. An Encyclopedia Article (1964)," *New German Critique* (3), 1974, pp. 49–55.

Hartley, Lauran R.

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



“Tibetan Publishing in the Early Post-Mao Period,” *Cahiers d’Extrême-Asie* 15, 2005, pp. 231–52. [doi:10.3406/asie.2005.1227](https://doi.org/10.3406/asie.2005.1227).

Ho, David Dahpon

“The Men Who Would Not Be Amban and the One Who Would. Four Frontline Officials and Qing Tibet Policy, 1905 -1911,” *Modern China* 34 (2), 2008, pp. 210–46.

Houn, Franklin W.

*To change a nation: Propaganda and indoctrination in communist China*. New York: Free Press of Glencoe, 1961.

Klu ma tshal Zla ba tshe ring

“Bod kyi tshags par gsar 'gyur gyi 'phel rim dang tshags par che chung gi rtsom rtsal la dpyad bsdur rags tsam byas pa.” [A Brief Comparative Analysis of the Development of Tibetan Newspaper Journalism and the Writing Techniques in large and small Newspapers]. *Bod ljongs slob grwa chen mo'i rig deb* 9 (2), 2001, pp. 60–69.

“Bod yig tshags par gyi 'phel rim dang mnam grangs de bzhin gnyer skyong bya tshul skor mdo tsam gleng ba.” [A Brief Discussion on the Development, Types, and Preservation State of Tibetan Language Newspapers]. *Bod ljongs slob grwa chen mo'i rig deb* (2), 2009, pp. 41–47.

Kobayashi, Ryosuke

“Zhang Yintang’s Military Reforms in 1906–1907 and their Aftermath. The Introduction of Militarism in Tibet,” *Revue d’Etudes Tibétaines* (53), 2020, pp. 303–40. Available online at [https://himalaya.socanth.cam.ac.uk/collections/journals/ret/pdf/ret\\_53\\_10.pdf](https://himalaya.socanth.cam.ac.uk/collections/journals/ret/pdf/ret_53_10.pdf)

Kolmaš, Josef

*Tibetan Books and Newspapers (Chinese collection): With Bibliographical Notes*. Asiatische Forschungen 62. Wiesbaden: Harrassowitz, 1978.

McLagan, Margaret Jane

"Mobilizing for Tibet: Transnational politics and diaspora culture in the post-cold war era." Ph.D, New York University, 1996.

Mgon po rdo rje

*Dus rabs gsar rnying gi bod: Smad cha* [Tibet of the Old and New Times: Second Volume]. Lo rgyus deb phreng 13. Dharamsala: Bod kyi dpe mdzod khang, 2015.

Moskaleva, Natalia N.

"Constructing the metanarrative of independence in *The Tibet Mirror* newspaper in the 1950s and 1960s." PhD, St. Petersburg State University, 2023.

Norbu, Tsewang

"Von medienscheuen Menschen zur virtuellen Gemeinschaft. Entwicklung der tibetischen Medienlandschaft im 20. Jahrhundert," *Tibetfocus* (112), 2011, pp. 4–6.

Pistorius, Kristin

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

Rgyal lo don grub

*The noodle maker of Kalimpong: My untold story of the struggle for Tibet*. Edited by Anne F. Thurston. London: Random House, 2015.

Roemer, Stephanie

*The Tibetan government-in-exile: Politics at large*. Routledge advances in South Asian studies 10. London: Routledge, 2008.

Roemer, Stephanie and Franz Xaver Erhard

"'Die Kanonen des Feindes gleichsam auf ihn selber richten'. Interkulturelle Auseinandersetzungen der Herrnhuter Brüder-

unität im Westhimalaja," *Zentralasiatische Studien* (36), 2007, pp. 237–62.

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." Diss. Heidelberg University Library, 2018. [doi:10.11588/heidok.00025156](https://doi.org/10.11588/heidok.00025156).

Schubert, Johannes

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

Sedláček, Kamil (ed.)

*Tibetan Newspaper Reader: In Two Volumes. 2 vols. Leipzig: Verlag Enzyklopädie, 1972.*

Shar ba thog med

"Nye rabs bod ljongs gsar 'gyur gyi byung ba mdo tsam brjod pa." [A Brief Account of the History of Modern Tibetan News]. *Bod-ljongs zhib-'jug* (69), 1999, pp. 60–70.

Sun Krung hran (孫中山 Sun Zhongshan 1866-1925) and Bstan pa lhun 'grub

*Rgyal yab sun krung hran mchog gis gsung ba: Dmangs gsum ring lugs (Gso skyid le'u gnyis) [三民主義 sanmin zhuyi Sun Yat-sen's Three Principles of the People]. Revised and enlarged ed. of the ed. Calcutta 1974. Ka ta: Krung dbyang gsar 'gyur par khang, 1985.*

Topgyal, Sonam

"The Tibet Mirror collection in LTWA, Dharamsala." In F. Wang-Toutain and M. Preziosi (eds.): *Cahiers du Mirror*, pp. 169–216, Paris: Collège de France, 2018.

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*, pp. 217–221, Paris: Collège de France, 2018.

Wang-Toutain, Françoise and Marie Preziosi

“Preface.” In Françoise Wang-Toutain and Marie Preziosi (eds.): *Cahiers du Mirror*, pp. v-viii, Paris: Collège de France, 2018.

Walravens, Hartmut

“The First Tibetan Serials,” *The Serials Librarian* 41(2), 2002, pp. 29–38. [doi:10.1300/J123v41n02\\_04](https://doi.org/10.1300/J123v41n02_04).

Walravens, Hartmut and I. Engelhardt (eds.)

*The First Tibetan Serial: August Hermann Francke's La-dvags-kyi-ag-bār (1904 - 1907); facsimile of a unique set in the archives of the Evangelische Brüderunität, Herrnhut. Neuerwerbungen der Ostasienabteilung / Staatsbibliothek zu Berlin Preussischer Kulturbesitz Sonderheft 22. Berlin: Staatsbibliothek, 2010.*

Xu Lihua 徐丽华

*Zang xue bao kan hui zhi* 藏学报刊汇志. [Tibetology Periodicals Collection Index]. Beijing: Zhongguo Zang xue chu ban she, 2003.

Zhou Decang 周德仓 (2005):

*Xizang xin wen chuan bo shi* 西藏新闻传播史. [History of News and Communication in Tibet]. Beijing: Zhong yang min zu da xue chu ban she, 2005.

## Appendix:

### *Systematic Description of the Divergent Discourses Corpus*

The DD Corpus contains the following datasets (1) the raw images of 17,115 newspaper pages; (2) enhanced images;<sup>35</sup> (3) e-texts, and bibliographic information for the 16 newspapers in mets xml-format. All datasets are stored in the Crossasia repository maintained by Staatsbibliothek zu Berlin once they become available.

### *Storage and File Format*

From scanning to text extraction, the project preserves a set of data at each stage. All datasets are stored using a human-readable file name in the following format encoding basic bibliographic information:

XXX\_YYYY\_MM\_DD\_ppp\_LL\_abcd  
Title\_Year\_Month\_Day\_Page\_Library\_Shelfmark  
TID\_1964\_01\_09\_001\_SB\_Zsn128162MR

TID\_1964\_01\_09\_001\_SB\_Zsn128162MR denotes page 1 of the January 9 issue of the year 1964 of Tibet Daily held at the Staatsbibliothek zu Berlin with shelfmark Zsn128162MR. This will ensure that at any time, the original page from which the digital copy was derived can be identified and located.

---

<sup>35</sup> See Sabbagh 2025 in this special issue, for a detailed description of how the project enhanced images for automatic text extraction.

## Newspaper Titles

#	Code	Newspaper title	Donors (shelfmark) <sup>36</sup>
1.	CWN	<i>Central Weekly News krung dbyang gsar 'gyur</i>	IT ( <a href="#">AC16810977</a> ); UW ( <a href="#">99133499060001452</a> );
2.	DTF	<i>Defend Tibet's Freedom rang dbang srung skyob gsar shog</i>	IT ( <a href="#">AC16810250</a> ); CU ( <a href="#">AN6.T6 .R36</a> ); BL
3.	FRD	<i>Freedom rang dbang gsar shog</i>	IT ( <a href="#">AC16809715</a> ); BD
4.	GDN	<i>Ganze Daily dkar mdzes nyin re'i gsar 'gyur</i>	OI ( <a href="#">XIV 92/1959</a> ); RB
5.	GOT	<i>Understanding go rtogs</i>	TS (not included corpus)
6.	GTN	<i>Gyantse News rgyal rtse gsar 'gyur</i>	LT
7.	KDN	<i>Kangding News dar mdo'i gsar 'gyur</i>	MV ( <a href="#">As Z Ag 10</a> )
8.	MJN	<i>Minjiang News min kyang tshags dpar</i>	MV ( <a href="#">As Z Ag 8</a> ); OI ( <a href="#">XIV 93/1959</a> )
9.	NIB	<i>News in Brief Gsar 'gyur mdor bsdus</i>	TL; MV ( <a href="#">As Z Ag 9</a> ); CU ( <a href="#">AN6.T6 G76</a> )
10.	QTN	<i>Qinghai Tibetan News Mtsho sngon bod yig gsar 'gyur</i>	CU ( <a href="#">AN6.T6M4</a> ); SB ( <a href="#">Zsn 128163 MR</a> ); MV ( <a href="#">As Z Ag 12</a> ); OI ( <a href="#">XIV 85/1959</a> )
11.	SGN	<i>South Gansu News kan lho gsar 'gyur</i>	OI ( <a href="#">XIV 90/1959</a> )
12.	TDP	<i>Tibet Daily Pictorial bod ljongs nyin re'i gsar 'gyur par ris</i>	OI ( <a href="#">XV 86/1959</a> )
13.	TID	<i>Tibet Daily bod ljongs nyin re'i gsar 'gyur</i>	OI ( <a href="#">XIV 91/1959</a> ); SB ( <a href="#">Zsn 128162 MR</a> ); IT ( <a href="#">AC16863326</a> )
14.	TIF	<i>Tibetan Freedom bod mi'i rang dbang</i>	IT ( <a href="#">AC16810977</a> )

<sup>36</sup> The shelfmarks link to respective online catalogues. Where no shelfmark is provided, the holdings are not catalogued.

#	Code	Newspaper title	Donors (shelfmark) <sup>36</sup>
15.	TIM	<i>Tibet Mirror</i> <i>yul phyog so so'i gsar 'gyur me long</i>	MV ( <a href="#">As Z Ag 11</a> ); IT ( <a href="#">AC16810250</a> ); CU ( <a href="#">DS786.A1 Y85</a> ); CF ( <a href="#">3 IET PER 1-28</a> )
16.	XNX	<i>South-West Institut for Nationalities</i> <i>Lho nub mi rigs slob grwa chen po</i>	OI ( <a href="#">XIV 89/1959</a> ); MV ( <a href="#">As Z Ag Z</a> )
17.	ZYX	<i>Central Institute for Nationalities</i> <i>Krung dbyang mi rigs slob grwa</i>	OI ( <a href="#">XIV 88/1959</a> )

### *Donor Institutions*

The newspapers in the DD Corpus were sourced from the following twelve libraries and private collections.

Code	Library	Newspapers
BD	Bodleian Library, Oxford, UK	FRD
BL	British Library, London, UK	DTF
CF	Collège de France, Bibliothèque des études tibétaines, Paris, France	TIM
CU	Columbia University Libraries, New York, USA	DTF; NIB; TIM; QTN
IT	Library of the Institute for South Asian, Tibetan, and Buddhist Studies, University Vienna	CWN; DTF; TIF; QTN; FRD;
MV	Grassi Museum für Völkerkunde, Leipzig, Germany	MJN; KDN; NIB; TIM; QTD; XNX; ZYX

<b>Code</b>	<b>Library</b>	<b>Newspapers</b>
OI	Oriental Institute (Czech Academy of Sciences), Prague, Czech Republic	MJN; TID; QTN; SGN;
RB	Robbie Barnett, London, UK (Private Collection)	GTN
SB	Staatsbibliothek zu Berlin, Germany	QTN; TID
TL	Library of Tibetan Works and Archives (LTWA), Dharamshala, India	NIB; FRD
TS	Tenzin Sonam, Dharamshala, India (Lhamo Tsering Archive/White Crane Films)	GTN
UW	University of Washington, Washington, USA	CWN






# Enhanced HTR Accuracy for Tibetan Historical Texts - Optimising Image Pre-processing for Improved Transcription Quality

Christina Sabbagh

(SOAS, University of London)

 Automatic transcription technology, or handwritten text recognition (HTR), unlocks new opportunities for analysing historical texts by transforming document images into machine-readable formats. This paper explores the development and evaluation of an image pre-processing pipeline to improve transcription accuracy for Tibetan historical newspapers. Images of historical texts often suffer from degradations introduced throughout their lifecycle, negatively affecting HTR accuracy. Additionally, the variability in image quality across archives poses challenges to creating a universally applicable pre-processing pipeline. This study compares three pre-processing pipelines, ultimately revealing a dynamic approach that could adapt to varying document conditions and that resulted in the highest transcription accuracy. This method offers a replicable solution for future research. We have also made the source code publicly available to support further exploration.

## 1 Introduction

Historical document preservation relies increasingly on digital transformation technologies. For Tibetan historical materials, this process confronts significant challenges: documents often suffer from

deterioration, including fading ink, paper damage, and inconsistent image quality. Such degradations impede handwritten text recognition (HTR),<sup>1</sup> which is essential for converting physical documents into searchable, analysable digital resources.

Current HTR technologies struggle with historical documents, particularly those with complex visual characteristics such as Tibetan newspapers in cases where the image quality is impaired. Manual image enhancement is time-consuming and impractical for large collections, necessitating automated solutions. Our research developed a pre-processing pipeline designed to improve image quality and the resulting transcription accuracy for these challenging historical documents.

This study addresses three primary questions:

- (1) Can targeted image pre-processing techniques effectively improve HTR transcription accuracy for degraded Tibetan historical documents such as newspapers?
- (2) How do different pre-processing approaches (traditional, deep learning, and hybrid) compare in addressing image quality challenges?
- (3) What methodology provides the most reliable and adaptable solution for transcribing historical Tibetan document images?

We compared three distinct image processing approaches: a traditional binarisation method, a deep learning-based technique, and a novel hybrid approach that dynamically selects the most appropriate method based on image quality. By evaluating these techniques, we

---

<sup>1</sup> While optical character recognition (OCR) refers to the translation of printed documents into machine-readable text, handwritten text recognition (HTR) is a similar process designed to handle the challenges of recognising text written in variable or inconsistent fonts, such as those created by handwriting or the differing fonts used across newspapers (Nockels *et al.* 2024: 149-150). HTR systems must account for these variations in style, size, and slant, which makes the process more complex than OCR.

aim to provide researchers with a robust, adaptable tool for digital document preservation.

By transforming deteriorating and rare documents into digital resources, this research offers an original approach to historical preservation by increasing the possibilities for large-scale analysis, broad accessibility, and long-term conservation of cultural heritage.

## 2 *Background*

### 2.1 *Dataset characteristics*

Experiments were conducted on a dataset provided by the “Divergent Discourses” project which applies digital philology methodologies to investigate the complex narratives of Tibetan history and identity emerging after the 1950 annexation of Tibet by the Chinese People’s Liberation Army. Central to this research is a collection of historical newspaper images curated from multiple archives, with handwritten text recognition (HTR) serving as a critical computational tool for transforming the document images into searchable digital resources.

As detailed by Erhard (2025) in this special issue, the Divergent Discourses Corpus of Tibetan-language newspapers currently comprises 16,718 pages from 16 newspapers, sourced from eight private collections and library archives across India, the United States, the United Kingdom, and Europe. Of these, 7,341 images are predominantly high-quality scans of original newspapers. In contrast, the majority – 9,377 images – were provided by the Staatsbibliothek zu Berlin (SB) and were digitised from microfilm. The microfilm, likely created in the late 1990s, was published by the China National Microforms Import & Export Corporation, Beijing, in the 2000s and presents several challenges. It includes missing and duplicated pages, as well as underexposed images with dark patches. Additionally, the collection preserved on microfilm shows signs of wear and water damage which poses significant difficulties for our use-case.

Drawing on the framework by Alaei *et al.* (2023), we analysed the document degradations visible within the dataset across three stages of a document's life cycle (Fig 1):

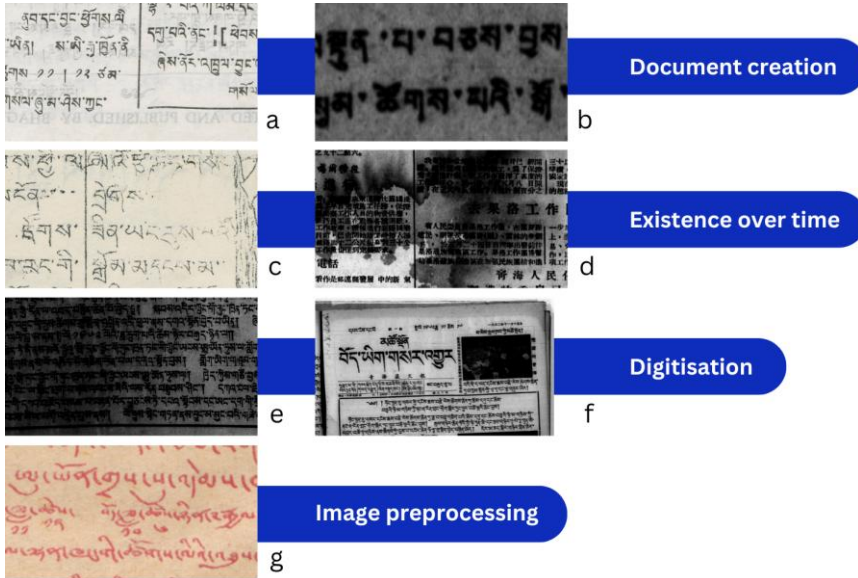


Figure 1 Visual examples of image degradation introduced at different stages within the document life cycle: a) bleed-through, b) blotchiness, c) fading, d) staining, e) uneven illumination, f) skewing, g) red text.<sup>2</sup>

- (1) Document Creation:
  - Bleed-through
  - Blotchy text, likely from low-quality paper or ink
- (2) Existence Through Time:
  - Text fading
  - Paper staining
- (3) Digitisation:
  - Uneven illumination during capture
  - Inconsistent page orientation
  - Compression and quality variations

<sup>2</sup> a) Tibetan Freedom, Nov. 1, 1965 p3; b) Tibet Daily, Jan. 1 1962 p1; c) Defend Tibet's Freedom, Aug. 20 1963 p3; d) Qinghai Tibetan News, Jun. 14 1955 p4; e) Qinghai Tibetan News, Jul. 5 1951 p2; f) Qinghai Tibetan News, Nov. 15 1952 p1; g) News in Brief, Dec. 1 1953 p1.

Red-text documents pose an additional challenge as they require distinct treatment during pre-processing compared to those containing black text. Failure to account for these chromatic differences generally resulted in less accurate HTR transcriptions, as traditional pre-processing approaches often reduced legibility of red text rather than improving it.

These cumulative degradations across the document life cycle obscure the textual features necessary for machine recognition, rendering standard HTR processes insufficient for reliable transcription.

## 2.2 *Image quality variation and representation bias*

Historical document collections inherently contain variability in quality arising from diverse archival sources, each reflecting distinct socioeconomic, geographical, and preservation contexts. This can result in variable HTR quality, introducing potential representation bias that could distort computational and historical analysis.

Ehrmann *et al.* (2023) illustrate this risk: a detected drop in word frequency during a historical period such as a war might not represent a genuine linguistic shift but could instead result from compromised paper quality. Newspapers from lower socio-economic strata may have been produced using lower-quality materials, increasing the likelihood of degradation over time and impacting HTR accuracy (Beelen *et al.* 2023). Such risks underscore the importance of effective image pre-processing.

Further, varying institutional digitisation strategies (Coutts 2016) can result in image quality differences. Research also suggests a correlation between digitisation quality and available financial resources (Smith & Cordell 2018). Institutions with limited economic and technical capabilities may produce lower-quality digital representations across different stages of the document life cycle. This variation in digitisation strategies can result in disproportionately poor-quality images within certain regions, potentially skewing research insights.

Image pre-processing techniques offer a methodological intervention, standardising transcription accuracy across heterogeneous document collections. By addressing image degradation introduced at each stage of a document's life cycle, researchers can minimise bias and enhance the reliability of computational and historical analyses.

This approach transforms technical challenges into an opportunity for more nuanced, comprehensive historical research, ensuring that marginal or less-preserved documents receive equal scholarly attention.

### 2.3 *Image binarisation*

Image binarisation is a frequently used pre-processing technique in digital document analysis that transforms complex images into black-and-white representations. By converting each pixel to either foreground (black) or background (white), binarisation helps isolate text and improve its machine-readability, particularly for historical documents with varying image qualities.

Researchers have developed three primary approaches to binarisation:

- **Uniform Thresholding:** Applies the same thresholding rule across the entire image, exemplified by the method developed by Otsu (1975).
- **Locally Adaptive Thresholding:** Tailors the threshold for each pixel based on local pixel neighbourhoods, making it particularly effective for images with uneven illumination. Sauvola (2000) and Niblack (1986) thresholding are examples of this approach.
- **Deep Learning-Based Methods:** Exemplified by the Berlin State Library's (SBB) binarisation model (Rezanezhad 2023), these methods use neural networks to make pixel-wise decisions based on both global and local image characteristics simultaneously.

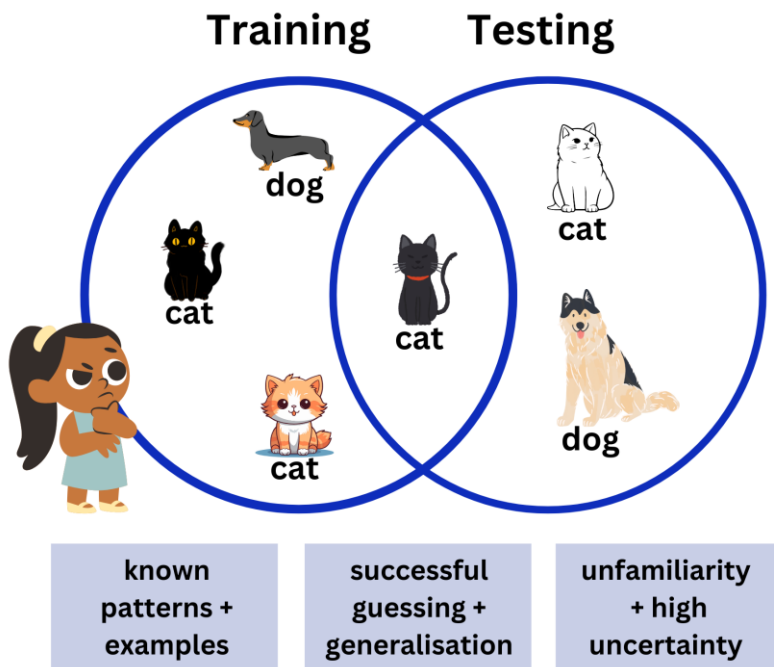


Figure 2 *A deep learning model learns to recognise patterns in data through training, much like how a child learns by studying examples. For instance, a parent may point to animals and label them as ‘dog’ or ‘cat.’ If the child only encounters small dogs and black or ginger cats during this learning phase, they become skilled at identifying those specific examples. However, in the real world, dogs vary in size, and cats can have different colours. When the child encounters a black cat they haven’t seen before, they can still correctly identify it. In contrast, encountering a large dog or a white cat may lead to misclassification. The more diverse the examples the child is exposed to, the better they become at generalising and accurately identifying unfamiliar animals.*

Different studies have yielded conflicting conclusions about which of these approaches best enhances HTR accuracy on document images. Some researchers found Otsu’s method superior for historical documents (Gupta *et al.* 2007; Rawat *et al.* 2021; Taş & Müngen 2021), while others demonstrated the effectiveness of locally adaptive techniques such as “Niblack Thresholding” (Jacsont & Leblanc 2023). These discrepancies likely arise from variations in underlying HTR models and the heterogeneous nature of document datasets.

More recently developed deep learning approaches, while promising, are not without limitations. These methods rely on training datasets to 'learn' patterns and make decisions, much like how humans learn by studying examples (Fig. 2). For instance, if a deep learning model is trained on a set of high-quality, clear document images, it becomes skilled at processing documents that look similar to those it has seen before. However, this process also means that the model might struggle with documents that look very different from its training examples, such as historical newspapers with stains, faded ink, or unusual layouts.

A common workaround is to use synthetic training data - artificially generated images designed to mimic real documents. While this is helpful for creating large datasets, it does not always prepare the model to handle the messy, unpredictable nature of real-world documents (Zhou *et al.* 2023). In contrast, traditional algorithmic methods do not rely on training data in the same way. Instead, they apply fixed rules and processes, which can make their performance more consistent across diverse document types and conditions.

These traditional methods often perform just as well as – or even better than – deep learning methods for certain tasks (Lins *et al.* 2021). This is because deep learning methods depend on their training datasets, which qualitatively align with the documents researchers plan to process. If the training data does not represent the target documents well, the model may fail to generalise effectively and produce less accurate results.

#### 2.4 *Document image pre-processing challenges*

The complex degradation-based challenges associated with historical document digitisation (outlined in Section 2.1) often require more than a single-step approach. Several research projects have developed tailored pre-processing strategies to address the degradation-based challenges.

For instance, Griffiths (2024) adjusted image sharpness, resolution, and noise levels to enhance a Tibetan manuscript dataset. Luo & van



der Kuijp (2024) implemented treatments including rotation correction, border removal, and contrast enhancement for Tibetan books and manuscripts. Rawat *et al.* (2021) demonstrated a multi-step approach for Garhwali textbook pages, utilising greyscaling, binarisation, morphological operations (described in Section 3.1.2), and skew correction.

However, not all pre-processing techniques are universally applicable. Some steps, such as border removal, may be redundant with advanced HTR tools such as Transkribus,<sup>3</sup> which can detect text regions without image trimming. Moreover, the lack of open-source methodologies has hindered the reproducibility and adaptability of previously used pre-processing pipelines. Finally, Jacsont *et al.* (2023) cautioned that combining multiple pre-processing treatments might produce lower-quality images than applying a single treatment, underscoring the importance of quantitative evaluation of pre-processing methods.

A limitation in current approaches is the assumption of uniform image quality. While researchers acknowledge that image qualities and required treatments can vary considerably within a single collection (Rawat *et al.* 2021), few have proposed systematic methods to address this variability. Our study addresses this gap by developing an adaptive pre-processing method that can accommodate diverse document characteristics.

Practical constraints further complicate the implementation of advanced pre-processing techniques. Many deep learning approaches (Anvari & Athitsos 2021; Zhou *et al.* 2023) require substantial computational resources and technical expertise to employ, making them challenging for projects with limited resources. Consequently, we prioritised methods that were well-established, openly accessible, and compatible with standard computing systems. The only exception was the deep learning-based binarisation approach which can still be

---

<sup>3</sup> Transkribus is a web-based platform which offers tools for the digitisation, text recognition, transcription and searching of historical documents. The Divergent Discourses project has used it to develop handwritten text recognition (HTR) models for the automatic transcription of its historical newspaper corpus.

run on standard computing systems, but more slowly than with sophisticated hardware.

Python libraries<sup>4</sup> such as OpenCV (Bradski 2000) and scikit-image (van der Walt 2014) offer a robust suite of pre-processing tools that are both extensively tested and accessible. By leveraging these resources, we aimed to develop a practical, reproducible methodology for researchers facing technological constraints.

Importantly, the complexity of pre-processing goes beyond mere technical optimisation. For images of historical documents - particularly those in lesser-studied languages - each image represents a potential trove of unique information. Inappropriate pre-processing can degrade image quality (Jacson & Leblanc 2023), or parts of the image, risking the loss of valuable historical insights. In the Divergent Discourses Corpus, many page images are unique, with no alternative copies available if images feature degradations. This highlights the importance of carefully considered and optimised pre-processing techniques.

## 2.5 *Image quality assessment*

Image quality assessment (IQA) is a technique for quantitatively evaluating document image characteristics, addressing the challenge of determining an image's legibility. Traditionally, image quality has been assessed through two primary approaches: Human-performed perceptual evaluation (subjective) and computational ('objective') models quantifying various forms of image degradation.

These assessment methods are broadly categorised into two types: no-reference and full-reference approaches. No-reference models, such as the one we employed, operate without an ideal-quality comparison image—important when working with unique historical

---

<sup>4</sup> A Python 'module' is a single file containing reusable code such as functions. A 'package' is a collection of related modules organised into folders. The term 'library' is generally used to describe collections of modules and packages that provide tools to perform a task, but it can also refer to a single module or package.

documents where perfect originals may not exist. Full-reference methods require an optimal version of the image for direct comparison.

In our research, we used the MANIQA model (Yang *et al.* 2022), a no-reference IQA tool. While initially developed and trained on a dataset of everyday photographs from the KONIQ-10k dataset (Hosu *et al.* 2020)—which predominantly features natural scenes, portraits, and urban environments—we found its quality assessment capabilities to be applicable to historical document analysis.

The KONIQ-10k dataset features images scored by humans based on quality indicators including noise (random colour or brightness variations that obscure details, such as graininess), compression artefacts, blur, exposure issues, and colour-related distortions. As several of these were degradations which appeared to affect HTR quality, the model demonstrated a reasonable degree of reliability in assessing document image readability across our historical Tibetan newspaper collection.

We systematically tested 18 IQA models, ultimately selecting MANIQA for its superior performance in reflecting our own transcription accuracy-based quality assessments (Appendix A).

### 3 *Experimental setup*

The objectives of our research were to evaluate whether targeted image pre-processing could improve handwritten text recognition (HTR) transcription accuracy for our Tibetan historical newspaper collection, and to identify which preprocessing steps improved recognition accuracy most effectively. We developed three pre-processing methods, comparing their performance against a baseline to identify the most effective image treatment strategies.

### 3.1 *Image pre-processing methods*

#### 3.1.1 *Baseline method*

Our baseline approach represented our minimal agreed requirements for document digitisation, including meeting Transkribus upload requirements as of June 2024. We prepared images by converting them to JPEG format, ensuring a minimum of 2500 pixels in at least one dimension, and maintaining a file size under 10 MB. This method simulated the most basic approach researchers might adopt when digitising historical documents, leaving images lightly pre-processed, providing a reference point for evaluating more sophisticated pre-processing techniques.

#### 3.1.2 *Foundational pre-processing pipeline*

We next developed our foundational pre-processing pipeline, deciding which treatments to include. Drawing on cautionary findings of Jacsont & Leblanc (2023) about the potential risks of combining multiple pre-processing treatments, we first isolated and evaluated individual image enhancement techniques.

We explored several pre-processing treatments (Fig. 3):

- **Greyscaling:** This treatment reduces the computational complexity of subsequent steps. To minimise processing time, we converted images to greyscale before testing the 'isolated' effects of each treatment.
- **Unsharp masking:** This method effectively sharpens character edges but simultaneously accentuates non-textual image features such as fold marks and stains. Our experiments revealed that unsharp masking can introduce additional

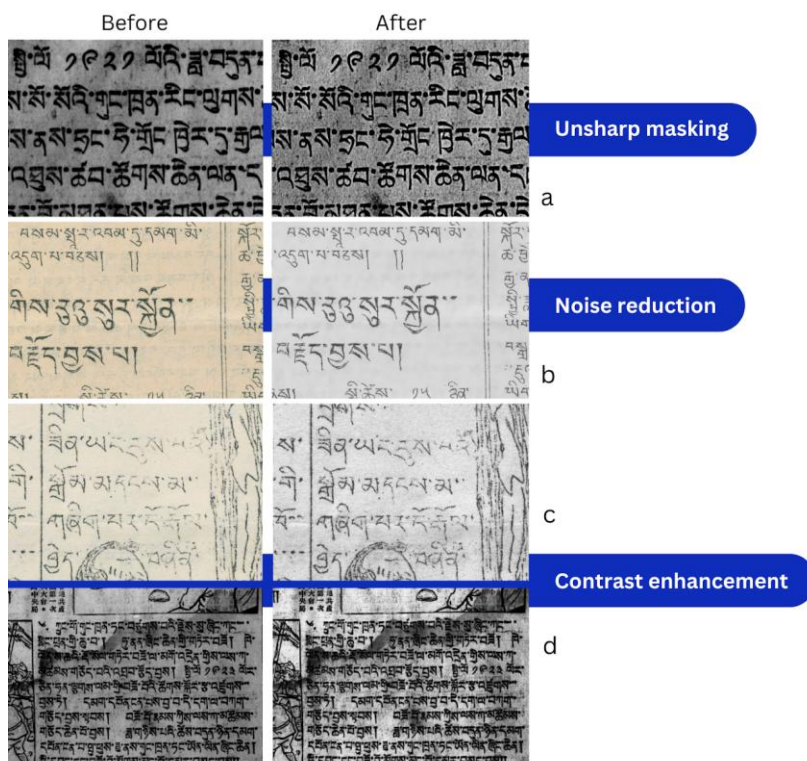


Figure 3 Images before and after having unsharp masking (a), noise reduction (b) and contrast enhancement (c, d) pre-processing treatments applied to them.<sup>5</sup>

speckling and background noise when combined with other treatments. We hypothesised that these artefacts would likely increase HTR transcription errors, as the model might misinterpret the enhanced noise as characters. Consequently, we excluded unsharp masking from our final pre-processing pipeline.

- **Noise reduction:** Using so-called fast non-local means denoising,<sup>6</sup> we effectively minimised digital artifacts such as

<sup>5</sup> a) Qinghai Tibetan News, July 5, 1952, p.4; b) Defend Tibet's Freedom, Aug. 20, 1963, p.8; c) Defend Tibet's Freedom, Aug. 20, 1963, p.3; d) Qinghai Tibetan News, July 5, 1952, p.4.

<sup>6</sup> Fast non-local means denoising aims to remove unwanted noise, such as graininess, from an image while preserving meaningful details. It does this by

page stains, bleed-through, and digitisation-introduced speckling. This technique preserved fine textual details while reducing background noise that could confuse HTR algorithms. We additionally experimented with: gaussian blur and median blur denoising (Bradski 2000). Gaussian and median blur denoising did not make a noticeable difference to the images so we did not progress with these.

- **Contrast enhancement:** To enhance the definition of characters, particularly faint and originally red text, we tested several contrast enhancement techniques. These included methods called: histogram equalisation, adaptive histogram equalisation (CLAHE), and a combination of normalisation (or 'contrast stretching') followed by CLAHE.

Histogram equalisation generally improved the human legibility of darker images but was less effective for lighter images, limiting its utility given the diverse lighting conditions in our dataset. CLAHE proved more effective, enhancing character definition across a wider range of image types. Combining normalisation with CLAHE yielded similar improvements in character definition while further increasing the contrast between originally red text and its background.

This combination marginally enhanced faint text and significantly improved the visibility of red text. However, it also exaggerated existing degradations such as staining and speckling, even after noise reduction. Given these drawbacks, we chose not to include contrast enhancement in our pipeline. Nonetheless, researchers working with less noisy datasets may find this approach beneficial for improving HTR results.

- **Binarisation (thresholding):** This represented our most nuanced intervention. We experimented with multiple thresholding approaches, ultimately finding that Sauvola and deep learning-based SBB binarisation provided the most significant

---

identifying similar patches throughout an image and averaging their values to smooth out the noise.

improvements in human legibility (Fig. 4).



Figure 4 A visual comparison of baseline (lightly treated) images, their counterparts treated with Sauvola and SBB binarisation<sup>7</sup>

<sup>7</sup> Tibet Daily, Jan. 1, 1962, p1 (top); Qinghai Tibetan News, Jun. 21, 1955, p1 (bottom)

We additionally experimented with the Otsu method, mean adaptive binarisation, and Niblack thresholding.

- **Dilation or erosion (morphological operations):** These strategies addressed image characteristics such as faint or blotchy characters (Fig. 5). Dilation increased the definition of faint characters. However, it dilated already blotchy characters, rendering them illegible, and amplified speckling and staining. Erosion intended to narrow blotchy characters often eroded the majority of the text, with no optimal setting that could simultaneously reduce blotchiness while preserving information in faint text. Given these limitations, we did not incorporate dilation or erosion into our final pre-processing pipeline.

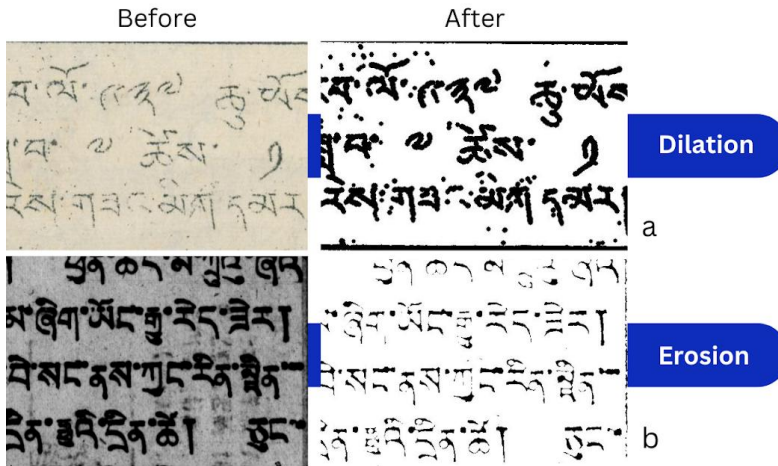


Figure 5 Images before and after dilation and erosion pre-processing operations have been applied to them.<sup>8</sup>

- **Skew correction:** This method addressed rotational irregularities. A projection profiling-based method (Reddy, 2019) appeared to prove most effective for our historical newspaper images. We also experimented with Brunner's (2024) "deskew" Python library, an implementation of the Hough Transform (Panzer, 2017), and the "skew\_correction" Python

<sup>8</sup> a) Defend Tibet's Freedom, Aug. 20, 1963, p.1; b) Qinghai Tibetan News, July 12, 1952, p.3.



library by Bhattarai (2019).

Rawat *et al.* (2021) suggested that super-resolution, used to enhance the resolution of images, was effective for low-resolution images but reduced the HTR accuracy for higher-resolution images. As we did not identify any low-resolution images (approximately 500 x 900 pixels) in our dataset, we did not experiment with super-resolution.

These experiments made the simplifying assumption that the treatments which resulted in the most improved human legibility would also result in the most improvement to machine legibility. Following isolated experiments, we combined the most promising of the above-listed treatments in a pipeline, resulting in the foundational pipeline outlined in Figure 6. The pipeline greyscaled the images, performed fast non-local means denoising, binarisation (either Sauvola binarisation or the deep learning-based SBB binarisation depending on the method), skew correction and compression to ensure the image was 10 MB or less.

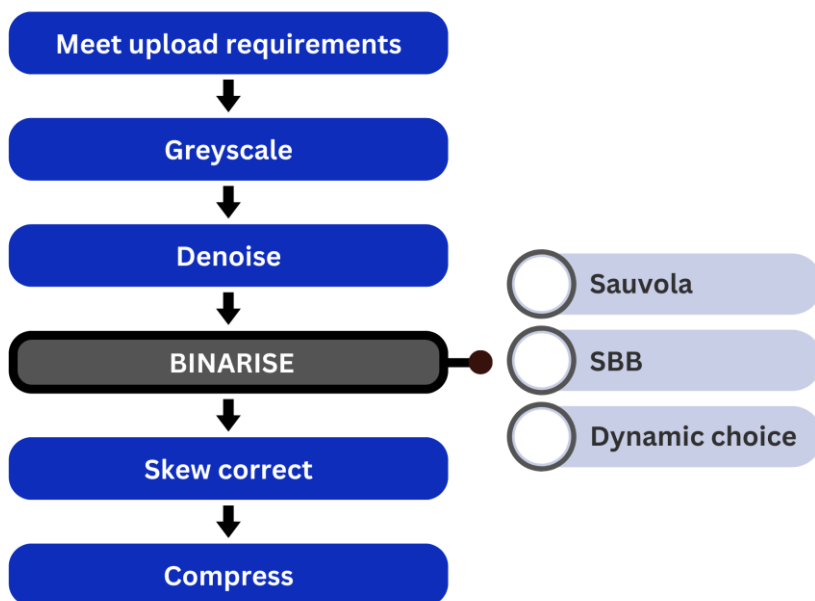


Figure 6 The foundational pipeline used for our three proposed methods. Each method uses a different binarisation method but otherwise shares the same pipeline.

For evaluation, we chose not to retrain our HTR model using the pre-processed images produced in our experiments. Instead, we evaluated pre-processing effects by inputting treated images into a model originally trained on untreated images. We hypothesised that this approach would enhance the model's robustness in dealing with image variation. In subsequent model iterations following our research, we incorporated both untreated and pre-processed images into our training set to enhance the model's robustness and ability to effectively generalise its learnings to unseen images by exposing the model to further variation.

### 3.1.3 *Method one: Sauvola binarisation pipeline*

The first approach in our pipeline (Fig. 6) employed Sauvola binarisation, a locally adaptive thresholding method particularly effective for addressing uneven illumination in historical images (Fig. 4, bottom). Unlike deep learning methods, which rely on diverse and representative training datasets, Sauvola binarisation applies fixed mathematical rules, producing consistent and predictable results regardless of the input image characteristics. This consistency made it a reliable choice for our dataset, despite its limitations.

Two settings, or hyperparameters, control Sauvola binarisation: window size and k-value. These parameters significantly influence the quality of binarisation, often requiring trade-offs between image subsets (Fig. 7).

For example, a larger window size enhances character definition but increases speckling in degraded images, while a higher k-value reduces noise but compromises contrast, especially in images with red text.

We tested k-values between 0.034 and 0.24, and window sizes from 11 to 41, ultimately selecting 0.14 for k-value and 21 for window size. This configuration maximised HTR accuracy across the dataset while minimising adverse effects on poor-quality and red-text images.

Despite these optimisations, red-text images remained particularly challenging, as boosting their contrast often degraded the quality of other subsets. The code for this method is publicly available (Sabbagh *et al.* 2024a).<sup>9</sup>

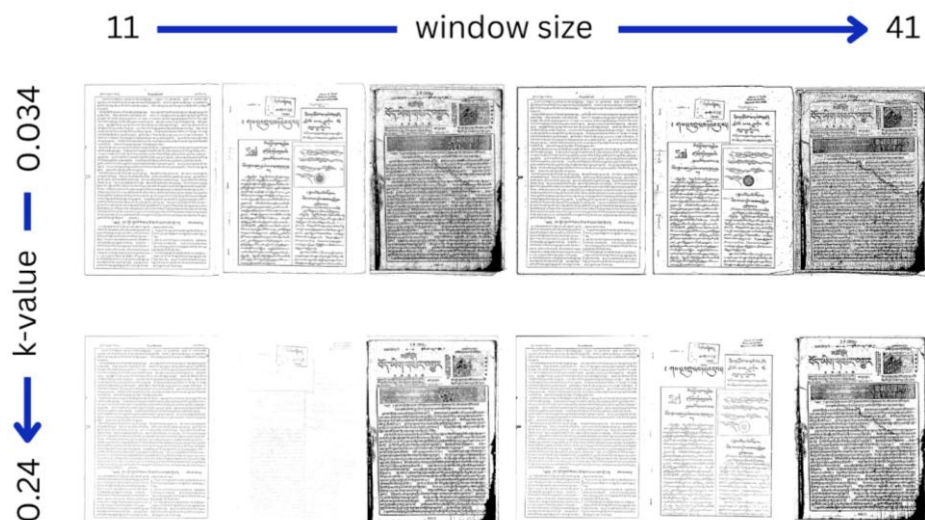


Figure 7 The trade-offs of using different Sauvola hyperparameter values (k-value, window size) in terms of image quality. The left-hand image of each trio is fair-quality, the middle image contains red text, and the right-hand image is poor-quality.<sup>1</sup> Each setting affects each image quality differently.

### 3.1.4 Method two: Deep learning binarisation pipeline (SBB binarisation)

The second method integrated the SBB deep learning-based binarisation model into our pipeline (Fig. 6). Unlike Sauvola binarisation, SBB employs a neural network to classify each pixel as foreground (text) or background by analysing patterns learned from training datasets, including those used in Document Image Binarization Contest (DIBCO) competitions, the Palm Leaf dataset

<sup>9</sup> Available online at [www.github.com/Divergent-Discourses/dd\\_preprocess](https://www.github.com/Divergent-Discourses/dd_preprocess) (accessed December 10, 2024).

(Burie *et al.* 2016), the Persian Heritage Image Binarization Competition (PHIBC) dataset (Ayatollahi & Nafchi 2013), and additional documents from the Berlin State Library (Rezanezhad 2023).<sup>10</sup>

The model analyses complex patterns, considering a pixel's immediate surroundings and its position within the broader context of the image when deciding whether it should be black or white. Since our Sauvola experiments (Section 3.1.3) indicated that tailoring binarisation settings by image type could improve results, SBB's adaptability offered a significant advantage. We anticipated that SBB would outperform traditional thresholding methods in handling complex challenges such as faded text, uneven illumination, and coloured or stained backgrounds.

### 3.1.5 *Method three: Forked binarisation pipeline*

Building on our evaluation of methods one and two, we developed a forked binarisation pipeline (Fig. 8) to combine their strengths and address the varying image qualities in our dataset. This approach aimed to improve HTR accuracy by dynamically selecting the most suitable binarisation method for each image. The code for this method is publicly available (Sabbagh *et al.* 2024b).<sup>11</sup>

Our analysis (Section 4) revealed distinct strengths for each method: Sauvola binarisation (method one) performed better on poor-quality images and those with red text, while SBB binarisation (method two) excelled with fair-quality images. However, some images achieved the highest HTR accuracy when left in their baseline, lightly processed state. Identifying these patterns motivated the creation of a pipeline capable of making automated decisions based on image quality. We therefore turned to image quality assessment (IQA) to identify

---

<sup>10</sup> More specifically, the model employs a hybrid CNN-Transformer architecture using a ResNet50-UNet encoder-decoder.

<sup>11</sup> Available at [www.github.com/Divergent-Discourses/dd\\_custom\\_preprocess](https://www.github.com/Divergent-Discourses/dd_custom_preprocess) (accessed December 10, 2024).

whether an image was fair-quality or poor-quality and pre-process this image using the pipeline best suited to its quality.

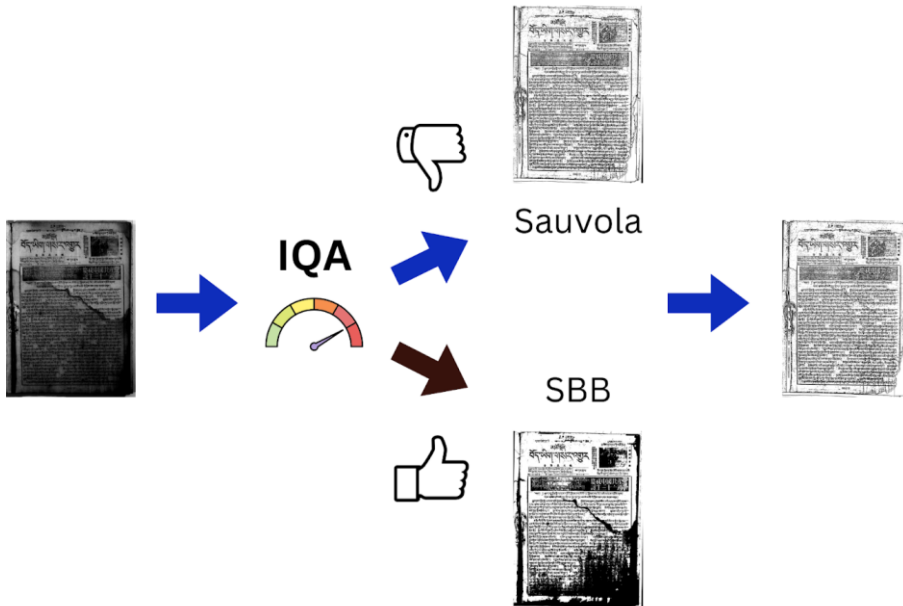


Figure 8 The forked binarisation pipeline and the potential path for one image<sup>1</sup> (blue arrows). An image quality assessment (IQA) method quality-scores an image. The image is considered poor-quality so the pipeline pre-processes it with the Sauvola binarisation-based pipeline. A score indicating good quality would have resulted in SBB binarisation-based pre-processing.

We used the MANIQA model (Yang *et al.* 2022), implemented via the PYIQA Python package (Chen & Mo 2021), to assign a perceptual quality score to each image, reflecting how a human might evaluate its visual quality. Images scoring above a threshold were classified as fair-quality and processed with SBB binarisation, while those scoring below the threshold were classified as poor-quality and processed with Sauvola binarisation. Appendix A details the rationale behind selecting MANIQA from among 18 tested IQA methods.

While this approach successfully classified images into two bins (fair- and poor-quality), it was unable to identify images that were better left in their baseline, lightly treated state. This issue stemmed from the binary nature of threshold-based classification, which

inherently divides images into only two classes: those scoring above or below the set threshold. Additionally, there were no consistently discernible visual characteristics or subjective criteria that reliably indicated when baseline, lightly treated images would outperform re-processed ones.

To mitigate this limitation, we decided to pre-process only a subset of the dataset—images from the library, Staatsbibliothek zu Berlin, comprising 9,377 of the 16,718 images in our total dataset. This subset predominantly consisted of poor-quality images<sup>12</sup> likely to benefit from pre-processing, along with a smaller number of good-quality images that appeared well-suited to SBB binarisation. It did not contain images with red text. Importantly, the subset appeared to contain few images that would have been most accurately transcribed in their baseline, lightly treated state, thereby minimising the risk of pre-processing inadvertently degrading transcription accuracy.

To adapt our approach to suit this library subset, we therefore adjusted Sauvola binarisation hyperparameters to better suit images without red text, setting the *k*-value to 0.24 and window size to 11. These settings improved binarisation quality for the remaining pages but were not ideal for the entire dataset, as discussed in Section 3.1.3. Using the forked pipeline, we applied Sauvola binarisation to poor-quality images and SBB binarisation to fair-quality images, predicting that this mixed approach would yield higher overall transcription accuracy.

Combining the forked pipeline evaluation with the new Sauvola hyperparameter settings presented a limitation: it was not possible to isolate the effects of the forked approach from those of the adjusted parameters. However, constraints within the Transkribus platform prevented us from conducting several separate evaluations. Despite this, the forked pipeline offers a pragmatic solution to the challenges posed by the varied quality of our dataset.

---

<sup>12</sup> These images are likely of poor quality due to the limitations of the original microfilm captures, produced by the China International Book Trading Corporation, Beijing, combined with the degradation of the microfilm itself over time.

### 3.2 *Evaluation methodology*

#### 3.2.1 *Test set composition*

To evaluate the methods described above, we curated a test set to assess HTR accuracy following pre-processing. The test set consisted of 86 images drawn from 11 newspapers, each manually transcribed to establish ground truth. This accounted for 0.5% of the total dataset. Although test sets typically represent a larger proportion of the overall data, expanding the test set was constrained by the time-intensive nature of manual transcription required to generate ground truth labels. A more standard 20% representation would have necessitated the transcription of approximately 3,400 images, which was not feasible within the scope of the project.

Instead, we prioritised diversity and ensured that the selected 86 images were representative of the key quality subsets identified within the dataset—specifically, poor-quality, fair-quality, and red-text images. Additionally, we targeted images exhibiting characteristics predicted to challenge HTR performance, such as bleed-through and underexposure. This approach allowed for a focused yet comprehensive evaluation of the model’s capabilities across the dataset’s most demanding scenarios.

To investigate how each method performed across the three main image quality categories in our dataset, we divided the test set into three subsets: fair-quality (42 images), poor-quality (29 images), and containing red text (15 images). These subsets were proportional to the estimated distribution of image qualities in the entire dataset, approximately 49%, 34%, and 17%, respectively. We did not use the test set images to train the HTR models.

Although we selected test set images to represent a wide range of characteristics hypothesised to contribute to HTR inaccuracies, the categorisation of images into the three subsets relied on subjective human judgment. These judgments may not have aligned perfectly with the features or patterns most relevant to deep learning algorithms. As a result, we may have placed some images in test subsets that did not align with how our HTR model interpreted them.

For example, an image the model struggled to transcribe may have been categorised as fair-quality based on human perception. This misalignment could lead to skewed subset results, potentially underestimating or overestimating the model's performance within specific categories. However, by ensuring that the selected images reflected diverse degradations, we aimed to mitigate this effect and capture the full spectrum of quality challenges.

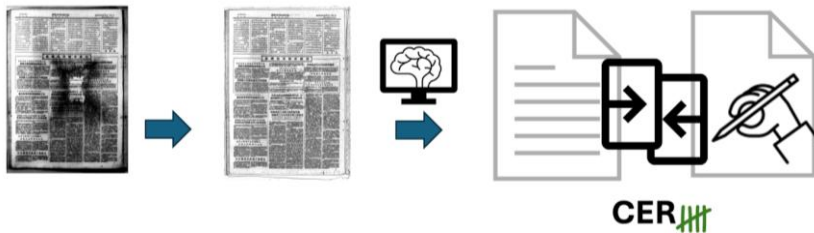
### 3.2.2 *Evaluating pre-processing methods*

To assess the performance of our three pre-processing methods, we executed the following (Fig. 9):

- (1) **Pre-processing:** For each method, we pre-processed the test set images, or for the baseline, ensured that the images met the minimum agreed requirements.
- (2) **HTR Processing:** The pre-processed images were inputted to the HTR model, which produced predicted transcriptions. Within Transkribus, we utilised two project-trained model prototypes: the field recognition model *TibNewsTR* and the handwritten text recognition (HTR) model *TibNewsOne4All 0.1*.
- (3) **CER Calculation:** For each image, we compared the predicted transcription with the ground truth to calculate the character error rate (CER) for each image (Appendix B). CER quantifies transcription errors - such as missing, incorrect, or additional characters - and ranges between 0 and 100, with higher scores indicating more errors. CER scores were computed using the Transkribus Expert client (Read-Coop SCE n.d.), as the standard Transkribus platform does not currently support this functionality (Transkribus n.d.).
- (4) **Averaging Results:** CER scores were averaged across subsets (fair-quality, poor-quality, red text) and across the entire test set to evaluate the effectiveness of each pre-processing method.



All CER scores were computed in a case-insensitive manner to avoid penalising transcription errors related to capitalisation, which did not make sense for Tibetan characters.



*Figure 9 Methodology to evaluate pre-processing methods. Page 4 of Qinghai Tibetan News, Jun. 14, 1955, is pre-processed with a given method (or left in its baseline, lightly processed state). The pre-processed image is inputted to the HTR model which outputs a predicted transcription. The predicted transcription is compared to the 'ground truth' transcription, outputting a character error rate (CER) value.*

During our evaluations, the Divergent Discourses project continued to develop new iterations of its HTR model. To ensure consistency, we selected a specific model iteration for all evaluations. Since then, the project has produced newer HTR model versions, trained on more diverse datasets and incorporating additional HTR steps (e.g., line polygon detection). As a result, our findings on the most effective pre-processing methods may not directly apply to the latest HTR models. This highlights a limitation of conducting evaluations alongside the ongoing development of interconnected components such as HTR models.

#### 4 Results

The experiments aimed to evaluate whether image pre-processing improved HTR transcription accuracy on our dataset of historical Tibetan newspaper pages. Additionally, we sought to determine which pre-processing method achieved the most accurate transcriptions overall, while considering that different image qualities might require distinct treatments.

This section first presents the visual outcomes of each pre-

processing method compared to the baseline. Next, it discusses transcription error rates for the entire test set, and finally examines transcription error rates for the subset of data selected for pre-processing based on the results of the first two pipelines.

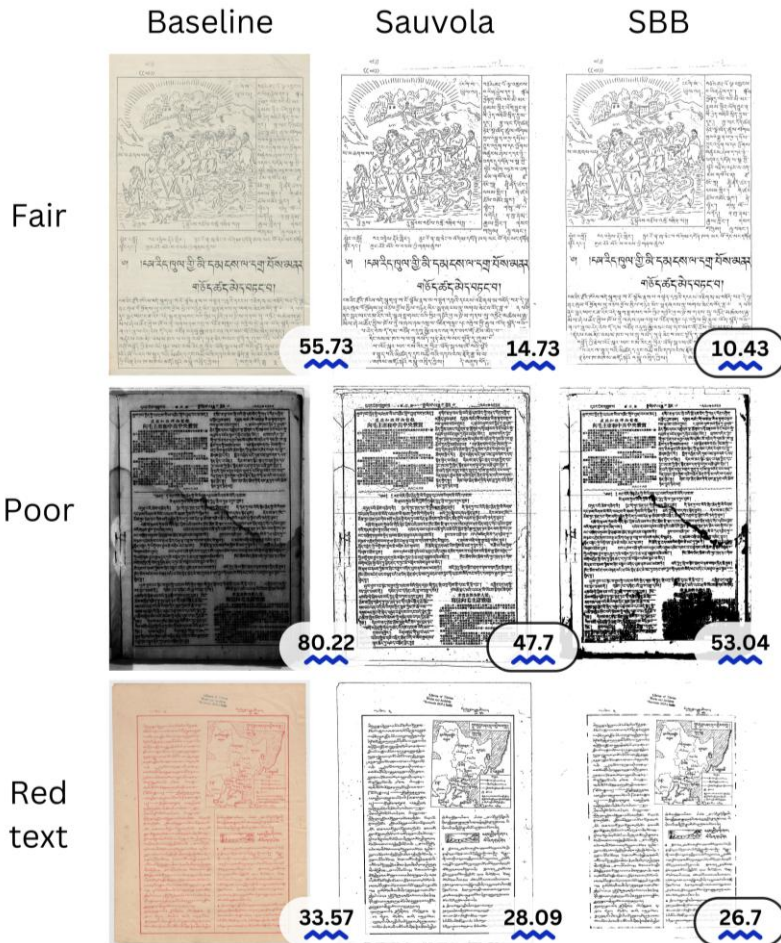


Figure 10 Visual outcomes and character error rate (CER) values when image is left in baseline, lightly treated state and after pre-processing using Sauvola and SBB binarisation-based pipelines. One image is shown for each test subset: fair-quality, poor-quality, and containing red text.<sup>13</sup>

<sup>13</sup> Defend Tibet's Freedom, Aug. 20, 1963, p.4 (fair-quality); Qinghai Tibetan News, Jul. 5, 1952, p.3 (poor-quality); News in Brief, Dec. 1, 1953, p.4 (red text).

#### 4.1 *Image pre-processing outcomes*

Visual outcomes of each pre-processing method, alongside baseline, lightly treated images, are shown in Figure 10. These examples illustrate how each method transformed the images prior to HTR processing.

#### 4.2 *Transcription error rates across test set*

Table 1 presents transcription error rates, measured using character error rate (CER), for each pre-processing method compared to the baseline. Median CER values suggest that all pre-processing methods improved HTR accuracy relative to the baseline. This trend was consistent with mean CER values, except for fair-quality images, where the mean for baseline images (38.89) slightly outperformed pre-processed results (39.86).

SBB binarisation yielded the lowest **median** CER for fair-quality images, suggesting that SBB binarisation was the most effective pre-processing method for fair-quality images. Fair-quality images left in their baseline, lightly pre-processed state had the lowest **mean** CER, suggesting that it was most effective to leave these images in their baseline, lightly treated state. While these findings are contradictory, median CER values and the resulting conclusions are likely to be more representative of performance, given the presence of outliers that skewed the data. This issue is explored further in Section 5.2. Sauvola binarisation was the most effective for poor-quality images and those containing red text, achieving the lowest CER values according to both mean and median values.

Across the entire test set, the forked binarisation method resulted in the lowest CER values, as hypothesised. Mean (39.91) and median (40.08) CER values for the forked method were both lower than those for any other method or the baseline, which ranged from 40.24 to 43.47 (mean) and 41.19 to 42.47 (median).

In general, transcription accuracy varied by image type, with red text images achieving the lowest CER scores (indicating the best

accuracy) across all methods, followed by fair-quality images, and finally poor-quality images.

*Table 1 Character error rates (CER) across test subsets (fair-quality, poor-quality, red text) and our entire test set. The best-performing method for each subset is shown in bold*

Pre-processing Method	Fair-quality (CER) ↓		Poor-quality (CER) ↓		Red text (CER) ↓		Overall (CER) ↓	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
Baseline	41.62	<b>38.89</b>	56.85	55.13	33.14	28.01	43.47	42.47
Method 1 (Sauvola binarisation)	43.06	44.42	<b>45.48</b>	<b>44.99</b>	<b>19.79</b>	<b>25.15</b>	40.24	41.25
Method 2 (SBB binarisation)	<b>39.86</b>	39.32	51.82	50.08	29.36	29.24	41.85	41.19
Method 3 (forked binarisation)	43.12	39.56	50.54	46.00	29.33	29.13	<b>40.08</b>	<b>39.91</b>

These results are illustrated in Figure 11, which shows the distribution of image-wise CER scores for each method. Baseline, lightly treated images generally had higher error rates – particularly for poor-quality images – with a wide spread of data points, reflecting the challenges faced by the HTR model in reliably identifying text in non-pre-processed images.

Both Sauvola binarisation and SBB binarisation achieved lower mean error rates compared to baseline, lightly treated images. Sauvola demonstrated greater improvement for poor-quality images, while SBB slightly outperformed for fair-quality images. Sauvola appeared to reduce the accuracy of some fair-quality images relative to leaving the images in their baseline, lightly processed state.

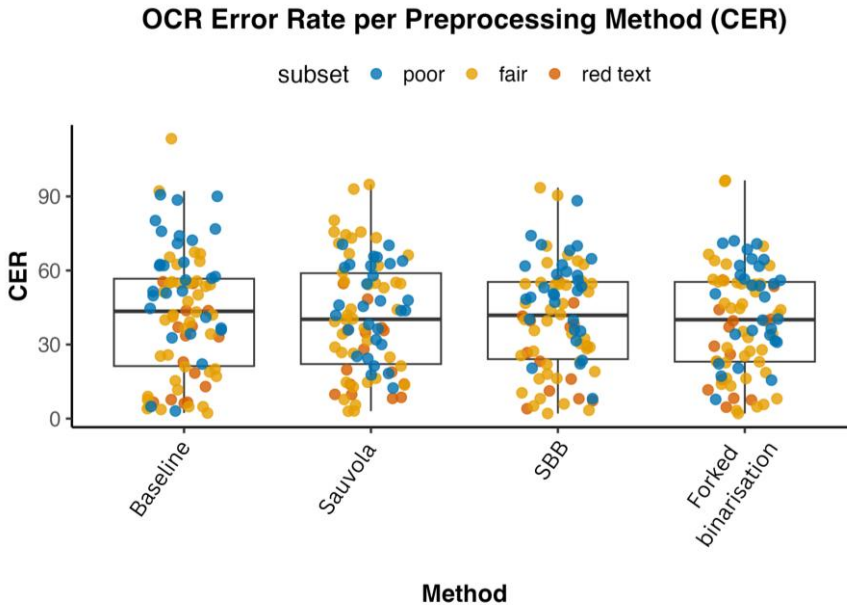


Figure 11 A box and whisker plot illustrating the distribution of individual image character error rates (CER) across methods. Values are shown for poor-quality (blue), fair-quality (orange) and red text (red) images.

The forked binarisation method showed the best overall performance, with the lowest mean error rate and tighter clustering of data points, indicating more consistent results across the test set. Furthermore, it produced fewer outliers with high error rates, reinforcing its effectiveness in preparing diverse image qualities for HTR.

### 4.3 Transcription error rates across library images

The test set included 27 images from the library, Staatsbibliothek zu Berlin, whose dataset we selected for pre-processing. Of these, 4 were fair-quality (15%), 23 were poor-quality (85%), and none contained red text. A qualitative review of the full dataset from this library suggests it contains a higher proportion of fair-quality images than represented in this test subset. Thus, while not strictly representative, this subset was evaluated to investigate whether the forked binarisation pipeline

would result in more accurate transcriptions than single-method approaches, as hypothesised.

Table 2 *Character error rates (CER) across images from the library, Staatsbibliothek zu Berlin, selected for pre-processing. Values are shown for test subsets (fair-quality, poor-quality) and overall. The best-performing method for each subset is highlighted in bold.*

Pre-processing Method	Fair-quality (CER) ↓		Poor-quality (CER) ↓		Overall (CER) ↓	
	Median	Mean	Median	Mean	Median	Mean
Baseline	41.62	43.66	61.90	61.87	56.85	59.17
Method 1 (Sauvola binarisation)	<b>20.56</b>	<b>26.01</b>	<b>47.70</b>	<b>48.03</b>	<b>45.98</b>	<b>44.77</b>
Method 2 (SBB binarisation)	22.67	28.89	53.50	55.01	53.04	51.14
Method 3 (forked binarisation)	21.77	27.74	53.95	49.51	53.79	46.28

Table 2 presents the results. Sauvola binarisation achieved the most accurate transcriptions for fair-quality, poor-quality, and the overall subset in terms of both mean and median CER values. All pre-processing methods improved transcription accuracy relative to the baseline, lightly treated images.

While the forked method (Method 3) was expected to outperform single-method approaches, the results showed Sauvola binarisation consistently achieved the lowest CER for all subsets.

## 5 Discussion

The results demonstrated that image pre-processing generally improved transcription accuracy compared to using baseline, lightly

treated images, with some exceptions for fair-quality images (Table 1, Table 2). However, the findings also highlighted that different image qualities benefited from distinct pre-processing treatments. This underscores the potential advantage of adaptive or multi-path pipelines that dynamically tailor pre-processing approaches based on detected image attributes. Notably, the forked binarisation pipeline delivered the highest overall transcription accuracy for a dataset containing images with diverse characteristics, validating its adaptive approach.

The SBB binarisation pipeline excelled in transcribing fair-quality images but struggled with poor-quality ones, performing worse than Sauvola binarisation. SBB's deep learning approach appeared to adapt well to the specific features of fair-quality images, producing cleaner backgrounds and more defined foreground text (Fig. 12).

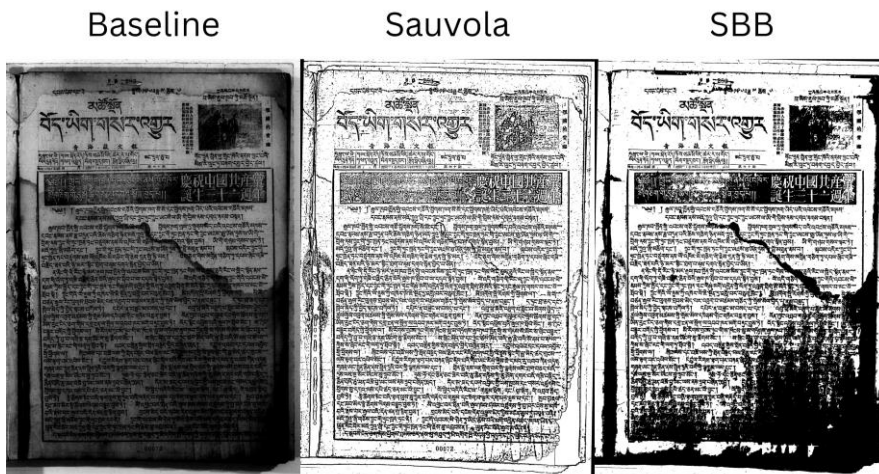


Figure 12 Visual effects of Sauvola and SBB binarisation compared to the baseline, lightly treated image. SBB binarisation resulted in cleaner, less speckled backgrounds but handled staining and dark patches poorly resulting in information loss.

However, it handled staining and dark patches poorly, often rendering these areas entirely black and causing significant information loss. This limitation is likely attributable to the training data for the SBB model, which may have predominantly consisted of images similar in quality to those classified as fair in our dataset. The model's lack of

exposure to poor-quality images during training likely limited its ability to address issues such as staining.

Sauvola binarisation, by contrast, was more effective for poor-quality images and those containing red text. Sauvola generally preserved more information in stained and distorted areas, minimising the risk of rendering significant portions of the page unusable. However, in some cases, SBB binarisation unexpectedly outperformed Sauvola for poor-quality images. These cases often involved minimal staining and significant bleed-through, where SBB's handling of the bleed-through yielded better results (Fig. 13).



Figure 13 *Sauvola binarisation used hyperparameters ( $k$ -value and window size) optimised for all three test subsets (fair-quality, poor-quality, red text) - the trade-off resulted in poorly binarised bleed-through relative to SBB binarisation (Qinghai Tibetan News, Jul. 12 1952 p2).*

Conversely, when low-resolution, blotchy characters were present, Sauvola occasionally outperformed SBB even for mostly fair-quality images. In such instances, SBB's tendency to merge stacked characters and increase blotchiness reduced transcription accuracy despite its ability to produce less speckled backgrounds (Fig. 14).





Figure 14 SBB binarisation sometimes resulted in character blotchiness rendering characters less legible than their Sauvola-binarised counterparts, even in images considered to be fair-quality (Tibet Daily, Jan. 1 1962, p1).

For red text, Sauvola binarisation generally resulted in superior transcription accuracy. This was despite SBB producing higher-contrast text, whereby achieving similar contrast with Sauvola would require fixed settings that would risk degrading pre-processing quality for other image types. Additionally, SBB often degraded page details, such as column borders, that may have served as useful clues for text region detection within the HTR pipeline. Sauvola preserved these details more consistently, contributing to its superior performance in these cases.

Interestingly, the baseline (lightly treated) inputs outperformed pre-processing for a subset of images, primarily from the fair-quality and red-text categories. Among these, 55% were fair-quality, a disproportionately high representation compared to the overall test set, while 24% were red text, also higher than expected. In contrast, only 21% were poor-quality, a lower proportion than their overall representation. These findings suggest that pre-processing is not universally beneficial, and its impact depends heavily on the specific characteristics of the input data.

For poor-quality images, pre-processing was particularly advantageous, likely due to its ability to correct distortions and enhance features which assisted HTR. For fair-quality images or those with specific challenges, such as red text, baseline (lightly treated) inputs sometimes retained details that pre-processing inadvertently degraded. For example, images with colour content risked reduced contrast when applying treatments optimised for other image types. This variability highlights the importance of selective or adaptable pre-processing approaches. Tailoring pre-processing to specific image attributes—such as quality, presence of red text, or colour—may mitigate issues of information loss and further optimise HTR performance.

### 5.1 *Forked binarisation method*

Recognising the benefits of tailoring pre-processing to specific image characteristics, our forked binarisation method dynamically selected the most suitable pre-processing approach to optimise transcription accuracy. This strategy aligns with the multi-pass approach described by Chastagnol (2013), who addressed the challenges of designing pipelines for heterogeneous datasets in a commercial setting.

Chastagnol's multi-pass algorithm applied several pre-processing methods to each image, quality-scored the results, and selected the highest-scoring version for HTR. In contrast, our approach is more computationally efficient: by quality-scoring images upfront, we determine the optimal pre-processing method, reducing the number of operations while still maintaining a focus on enhancing transcription accuracy.

On the overall test set, our approach's adaptability to image characteristics resulted in higher overall transcription accuracy compared to using a single pre-processing method. The image-wise evaluation data indicated that optimising hyperparameters for Sauvola binarisation further improved accuracy for pages where Sauvola was the preferred method. For images where the pipeline correctly selected Sauvola, the character error rate (CER) values were

generally lower than those achieved with our formerly-selected Sauvola hyperparameter values, except for on images containing red text. For red-text images, CER values were higher than expected due to the hyperparameters having been optimised for non red-text images. If the pipeline was enhanced to also leave images in their baseline (lightly treated) state when beneficial, transcription accuracy would likely improve further.

On the Staatsbibliothek zu Berlin image test set, the adaptability of the forked binarisation pipeline did not yield higher overall transcription accuracy. This is likely because 85% of the images were poor-quality, which would predominantly benefit from Sauvola binarisation based on our results. Given the relatively homogeneous nature of this test set, the dynamic approach offered by the forked pipeline was not necessary and generated errors. However, the library test set did not fully represent the broader image quality distribution from the library. A qualitative review suggests that the actual library image set includes a higher proportion of fair-quality images, which would likely benefit more from the dynamic selection provided by the forked binarisation pipeline.

A limitation of our approach was that the choice between Sauvola and SBB binarisation did not strictly correlate with the image quality categories (e.g., fair or poor). These quality labels served as proxies for the true objective: determining the optimal pre-processing method (either Sauvola or SBB) for a given image. The reasons why some images are more accurately transcribed with one method over the other remain speculative and are likely not strictly related to image quality. Ideally, a neural network could be trained to predict the most suitable pre-processing method for each image, effectively emulating the decision-making process of our approach. However, generating sufficient training data for this task was outside the scope of this project.

### 5.2 *Data outliers*

As discussed in Section 4.2, our dataset contained a small number of outliers that made the median CER values more reliable than the mean. These outliers included images that were automatically transcribed significantly more accurately in their baseline, lightly treated form than after applying any of the three pre-processing methods. We suspect these images were inadvertently included both in the training set for our HTR model and in the test set. As a result, the model likely performed better on the baseline (lightly treated) images because it had been trained on these exact baseline images and their corresponding transcriptions, essentially "memorising the answers" during training. In contrast, the pre-processed versions of these images, which the model had not encountered during training, were transcribed with greater error, highlighting the importance of proper dataset partitioning in machine learning model development.

Another source of outliers was pages containing a significant amount of Chinese text. Our HTR model had been trained primarily on Tibetan text with limited exposure to Chinese. Consequently, the transcription accuracy for these pages was lower than for others in the dataset. Despite these limitations, these pages containing Chinese language still enabled valid comparisons across pre-processing methods, as they were consistently included in all experiments.

### 5.3 *Future work*

Future work could explore more advanced approaches to tailoring pre-processing methods within the pipeline while remaining mindful of budgetary constraints typical of project-based work. One promising avenue is the use of clustering algorithms to automatically group images by quality, enabling the application of targeted pre-processing strategies to each cluster. Additionally, computer vision techniques could be employed to identify specific problem areas within images—such as staining, uneven illumination, or coloured text—and customise pipelines for individual images.

For example, the Turing Institute’s MapReader tool (Wood *et al.* 2024) classifies patches of an image and could potentially be adapted to identify visual features in historical newspaper images. This approach aligns with budgetary constraints, as it would require relatively simple patch-wise annotation of visual features rather than the resource-intensive task of manual transcription for training a model.

It is also important to recognise that achieving a CER of zero is rare; HTR transcriptions will almost always contain some degree of imperfection. Researchers must account for such errors—or ‘HTR noise’—when using these transcriptions in downstream applications, such as building databases or conducting computational textual analysis. For instance, the Impresso project addresses optical character recognition (OCR) noise by offering a keyword suggestion tool that proposes fuzzy matches for user search queries, ensuring that relevant texts are retrieved even when transcription errors occur (Düring *et al.* 2024). Similar tools could be developed to support Tibetan studies, where HTR challenges are compounded by the complexity of the script and the variability of historical document conditions.

## 6 Conclusion

Our research addressed the challenge of enhancing HTR transcription accuracy for historical Tibetan newspaper images through tailored and adaptive pre-processing strategies. By evaluating three distinct binarisation approaches, we demonstrated the important role of context- and quality-aware image enhancement techniques in improving HTR outcomes for heterogeneous document collections.

The forked binarisation method we developed offers a promising solution to the complexities of digitising heterogeneous historical texts. Unlike uniform pre-processing strategies, our approach dynamically selects the most appropriate method based on individual image characteristics, striking a balance between accuracy and computational efficiency.

Beyond the technical advancements, our work underscores the broader implications of HTR quality. Variations in document condition can introduce biases in the accessibility of historical texts, potentially affecting socio-political research and cultural preservation efforts. This highlights the importance of nuanced, adaptive digitisation approaches that consider the importance of technical performance in relation to evenly distributed access.

By openly sharing our method, we aim to support future research in Tibetan studies, document preservation, and the broader field of historical HTR. Refining these adaptive pre-processing strategies could significantly enhance access to historical texts that might otherwise remain inaccurately transcribed and overlooked, thereby supporting their continued study and preserving their cultural significance.

### Bibliography

- Alaei, Alireza, Vinh Bui, David Doermann, and Umapada Pal  
"Document Image Quality Assessment: A Survey," *ACM Computing Surveys*, 56 (2), 2023, pp. 1–36. [doi: 10.1145/3606692](https://doi.org/10.1145/3606692)
- Anvari, Zahra, and Vassilis Athitsos  
"A Survey on Deep Learning Based Document Image Enhancement," *arXiv preprint*, 2021. [doi:10.48550/arXiv.2112.02719](https://doi.org/10.48550/arXiv.2112.02719)
- Ayatollahi, Seyed Morteza, and Hossein Ziaei Nafchi  
"Persian heritage image binarization competition (PHIBC 2012)." In *First Iranian Conference on Pattern Recognition and Image Analysis (PRIA)*, IEEE, 2013, pp. 1-4. [doi:10.1109/PRIA.2013.6528442](https://doi.org/10.1109/PRIA.2013.6528442)
- Beelen, Kaspar, Jon Lawrence, Daniel C.S. Wilson, and David Beavan  
"Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan," *Digital Scholarship in the Humanities* 38 (1), 2023, pp. 1–22. [doi:10.1093/llc/fqac037](https://doi.org/10.1093/llc/fqac037)

Bhattacharai, Ashuta (username ashuta03)

"Automatic Skew Correction Using Corner Detectors and Homography," *GitHub repository*, 2019. Available online at [https://github.com/ashuta03/automatic skew correction using corner detectors and homography](https://github.com/ashuta03/automatic-skew-correction-using-corner-detectors-and-homography) (accessed July 3, 2024).

Bradski, Gary

"The OpenCV Library," *Dr. Dobb's Journal of Software Tools* 25 (11), 2000, pp. 120, 122-125.

Brunner, Stéphane

"Deskew: Skew detection and correction in images containing text," *Python Package Index (PyPI)*, 2024. Available online at <https://pypi.org/project/deskew/> (accessed July 3, 2024).

Burie, Jean-Christophe, Mickaël Coustaty, Setiawan Hadi, *et al.*

"ICFHR2016 competition on the analysis of handwritten text in images of Balinese palm leaf manuscripts." In *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, 2016, pp. 596-601. [doi:10.1109/ICFHR/2016.107](https://doi.org/10.1109/ICFHR/2016.107)

Chastagnol, François

"Building an Image Pre-processing Pipeline in Python," *YouTube video, Next Day Video*, 2013. Available online at <https://www.youtube.com/watch?v=B1d9dpqBDVA> (accessed October 29, 2024).

Chen, Chaofeng, and Jiadi Mo

"IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment," 2021. Available online at <https://github.com/chaofengc/IQA-PyTorch> (accessed October 29, 2024).

Coutts, Margaret

*Stepping Away from the Silos: Strategic Collaboration in Digitisation*. Chandos Publishing, 2016.

## DIBCO

"Datasets," Last updated October 4, 2023. Available online at <https://dib.cin.ufpe.br/#!/datasets> (accessed October 29, 2024).

Sabbagh, Christina, Franz Xaver Erhard, Robert Barnett, and Nahan W. Hill

"Divergent Discourses Custom Image Preprocessing (Sauvola Binarisation)," *Zenodo*, 2024a. [doi:10.5281/zenodo.14525692](https://doi.org/10.5281/zenodo.14525692).

"Divergent Discourses Custom Image Preprocessing (Forked Binarisation)," *Zenodo*, 2024b. [doi:10.5281/zenodo.14523007](https://doi.org/10.5281/zenodo.14523007).

Düring, Marten, Estelle Bunout, and Daniele Guido

"Transparent Generosity: Introducing the *impresso* Interface for the Exploration of Semantically Enriched Historical Newspapers," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2024, pp. 35–55. [doi:10.1080/01615440.2024.2344004](https://doi.org/10.1080/01615440.2024.2344004)

Erhard, Franz Xaver

"The Divergent Discourses Corpus: A Digital Collection of Early Tibetan Newspapers of the 1950s and 1960s," *Revue d'Études Tibétaines*, (74), 2025, pp. 45–81.

Ehrmann, Maud, Edouard Bunout, and Frédéric Clavert

"Digitised Historical Newspapers: A Changing Research Landscape." In *Newspapers—A New Eldorado for Historians*, 2023, pp. 1–22. [doi:10.1515/9783110729214-001](https://doi.org/10.1515/9783110729214-001)

Griffiths, Rachael

"Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script," *Revue d'Études Tibétaines*, (72), 2024, pp. 43–51. [doi:10.1553/TibSchol\\_ERC\\_HTR](https://doi.org/10.1553/TibSchol_ERC_HTR).

Gupta, Maya, R, Nathaniel P. Jacobson, and Erik K. Garcia

"OCR Binarization and Image Pre-Processing for Searching Historical Documents," *Pattern Recognition* 40 (2), 2007, pp. 389–397. [doi:10.1016/j.patcog.2006.04.043](https://doi.org/10.1016/j.patcog.2006.04.043).



- Hosu, Vlad, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe  
"KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment," *IEEE Transactions on Image Processing* 29, 2020, pp. 4041–4056. [doi:10.1109/TIP.2020.2967829](https://doi.org/10.1109/TIP.2020.2967829)
- Jacsont, Pauline, and Elina Leblanc  
"Impact of Image Enhancement Methods on Automatic Transcription Trainings with eScriptorium," *Journal of Data Mining & Digital Humanities*, 2023. [doi:10.46298/jdmdh.10262](https://doi.org/10.46298/jdmdh.10262)
- Lins, Rafael Dueire, Rodrigo B. Bernardino, Edward B. Smith, *et al.*  
"ICDAR 2021 Competition on Time-Quality Document Image Binarization." In *Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV, vol. 16*, Springer International Publishing, 2021, pp. 708–22. [doi:10.1007/978-3-030-86337-1\\_47](https://doi.org/10.1007/978-3-030-86337-1_47)
- 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'Études Tibétaines* (72), 2024, pp. 26–42. Available online at [https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret\\_72\\_02.pdf](https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_72_02.pdf) (accessed January 26, 25).
- Niblack, Wayne  
*An Introduction to Digital Image Processing*. Englewood Cliffs: Prentice-Hall, 1986.
- 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. 148–167. [doi:10.1108/JD-09-2023-0183](https://doi.org/10.1108/JD-09-2023-0183)
- Otsu, Nobuyuki  
"A Threshold Selection Method from Gray-Level Histograms," *Automatica* (11), 1975, pp. 23–27.

Panzer, Jason Drew

"Hough Transform implementation," *GitHub*, 2017. Available online at <https://gist.github.com/panzerama/beebb12a1f9f61e1a7aa8233791bc253> (accessed July 3, 2024).

Rawat, Sukhbindra Singh; Ashutosh Sharma, and Rachana Gusain

"Analysis of Image Pre-processing Techniques to Improve OCR of Garhwali Text Obtained Using the Hindi Tesseract Model," *ICTACT Journal on Image & Video Processing*, 12 (2), 2021. [doi:10.21917/ijivp.2021.0366](https://doi.org/10.21917/ijivp.2021.0366)

Read-Coop SCE

"Transkribus Expert Client, Version 1.28.0," Software. n.d. Available online at <https://readcoop.eu/transkribus/> (accessed November 3, 2024).

Reddy, Susmith

"Pre-Processing in OCR!!!" *Towards Data Science*, 2019. Available online at <https://towardsdatascience.com/pre-processing-in-ocr-fc231c6035a7>. (accessed July 3, 2024).

Rezanezhad, Vahid, Konstantin Baierer, and Clemens Neudecker

"A Hybrid CNN-Transformer Model for Historical Document Image Binarization," In Antonacopoulos, Apostolos, Christian Clausner, Maud Ehrmann, Kai Labusch, and Clemens Neudecker (eds.) *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing (HIP) 2023*, San José, 2023. [doi: 10.1145/3604951.3605508](https://doi.org/10.1145/3604951.3605508).

Sauvola, Jaakko, J. and Matti K. Pietikainen

"Adaptive Document Image Binarization," *Pattern Recognition* 33, 2000, pp. 225–236, [doi:10.1016/S0031-3203\(99\)00055-2](https://doi.org/10.1016/S0031-3203(99)00055-2).

Smith, David A., and Ryan Cordell

"A Research Agenda for Historical and Multilingual Optical Character Recognition," *NULab*, Northeastern University, 2018, pp. 36. [doi:10.1177/0961000611434760](https://doi.org/10.1177/0961000611434760).

Smith, Lucy, and Jennifer Rowley

"Digitisation of Local Heritage: Local Studies Collections and Digitisation in Public Libraries," *Journal of Librarianship and Information Science*, 44 (4), 2012, pp. 272–80.

Taş, İdal Çetin and Ahmet Anil Müngen

"Using Pre-Processing Methods to Improve OCR Performances of Digital Historical Documents." In *Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, 2021, pp. 1–5. [doi:10.1109/ASYU52992.2021.9598972](https://doi.org/10.1109/ASYU52992.2021.9598972).

Transkribus

"6. Computing Accuracy", n.d. Available online at <https://help.transkribus.org/computing-accuracy> (accessed December 7, 2024).

Wood, Rosie, Kasra Hosseini, Kalle Westerling, *et al.*

"MapReader: Open Software for the Visual Analysis of Maps," *Journal of Open Source Software* 9 (101), 2024, p. 6434. [doi:10.21105/joss.06434](https://doi.org/10.21105/joss.06434).

Yang, Sidi; Tianhe Wu, Shuwei Shi, Shanshan Lao, *et al.*

"Maniqa: Multi-Dimension Attention Network for No-Reference Image Quality Assessment." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1191–1200. [doi:10.1109/CVPRW56347.2022.00126](https://doi.org/10.1109/CVPRW56347.2022.00126)

Zhou, Yanxi, Shikai Zuo, Zhengxian Yang, Jinlong He, Jianwen Shi, and Rui Zhang

"A Review of Document Image Enhancement Based on Document Degradation Problem," *Applied Sciences*, 13 (13), 2023, p. 7855. [doi:10.3390/app13137855](https://doi.org/10.3390/app13137855).

## Appendices

### *Appendix A Selecting an image quality assessment (IQA) method*

To select the most effective Image Quality Assessment (IQA) method for use in our forked binarisation pipeline (Section 3.1.5), we conducted an experiment using the IQA methods available through the PYIQA Python package (Chen *et al.* 2021). Our goal was to determine which method would most accurately assess the quality of images in our test set.

We began by evaluating the performance of three methods; the first leaving the images in their baseline (lightly treated) state, the second applying the Sauvola binarisation method (Section 3.1.3), and the third applying the SBB binarisation method (Section 3.1.4). These evaluations provided image-wise Character Error Rates (CER) for each image, with three CER values per image - one for each method - indicating how accurately our HTR model transcribed each version of the image. Using these CER values, we classified each image according to the method that resulted in the most accurate transcription.

Since the IQA methods output a quality score for each image, we categorised the images into two classes based on a chosen threshold: 'fair-quality' and 'poor-quality'. Images most accurately transcribed by the Sauvola binarisation pipeline were labelled as 'poor-quality', while those best transcribed by the SBB binarisation pipeline were labelled as 'fair-quality'. For images which were most accurately transcribed when left in their baseline (lightly treated) state, we labelled them according to the second-best method.

Using each IQA method (listed in Table 3), we obtained quality scores for the images in our test set. We then calculated the mean quality score across the dataset and used this as the threshold to predict whether each image should be labelled as 'fair-quality' or 'poor-quality'. This process was repeated for all 18 methods, and we tracked the number of correct classifications across the entire dataset, as well as within each quality subset (fair-quality, poor-quality, red

text). This ensured that the methods did not perform disproportionately well on one subset and poorly on others.

Table 3 IQA model variants included in our experiments, alongside results over our entire test set. Names following dashes refer to the datasets upon which models were trained on.

Method Number	Method Name	Correctly predicted labels (/86)	Model Source
1	ARNIQA-clive	45	Agnolucci <i>et al.</i> 2024
2	ARNIQA-csiq	37	
3	ARNIQA-flive	39	
4	ARNIQA-kadid	50	
5	ARNIQA-koniq	42	
6	ARNIQA-live	42	
7	ARNIQA-spaq	42	
8	ARNIQA-tid	47	
9	BRISQUE	42	Mittal <i>et al.</i> 2012
10	CNNIQA	51	Kang <i>et al.</i> 2014
11	DBCNN	47	Zhang <i>et al.</i> 2018
12	HyperIQA	50	Su <i>et al.</i> , 2020
13	MANIQA-kadid	51	Yang <i>et al.</i> 2022
14	MANIQA-koniq	<b>53</b>	
15	MANIQA-pipal	52	
16	TReS-flive	33	Golestaneh <i>et al.</i> 2022
17	TReS -koniq	50	
18	WaDIQaM-nr	33	Bosse <i>et al.</i> 2017

Among the IQA methods tested, the quality scores from MANIQA-koniq resulted in the highest number of correctly classified images. As a result, we incorporated this model into our forked binarisation pipeline and adjusted the threshold from 0.2 to 0.335, which appeared to be more suitable for classifying our particular dataset.

*Appendix B Character error rate (CER)*

In this study, character error rate (CER) represents the percentage of characters which the HTR model has transcribed incorrectly, as determined through comparison with a 'ground truth', or reference, transcription. A lower CER value indicates a higher accuracy HTR model. It can be calculated as shown below.

N is the number of characters in the ground truth transcription. S is the number of character substitutions, D is the number of character deletions and I is the number of character insertions relative to the 'ground truth'.

$$CER = \frac{S + D + I}{N} \times 100$$

A character would be considered a substitution when incorrect but corresponding to one character in the ground truth (e.g. if the HTR model predicts the word ལྷལ but the correct transcription is ལྷལ།). A deletion would be where a character was missing from the predicted transcription (e.g. ལྷལ). An insertion would be where the HTR model incorrectly predicted an additional character.

**Supplementary bibliography**

Agnolucci, Lorenzo, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo

"Arniqa: Learning Distortion Manifold for Image Quality Assessment." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 189–198 [doi:10.1109/WACV57701.2024.00026](https://doi.org/10.1109/WACV57701.2024.00026)

Bosse, Sebastian, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek

"Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," *IEEE Transactions on Image Processing* 27 (1), 2017, pp. 206–219. [doi:10.1109/TIP.2017.2760518](https://doi.org/10.1109/TIP.2017.2760518)


- Golestaneh, S. Alireza, Saba Dadsetan, and Kris M. Kitani  
“No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency.” In *IEEE Computer Society*, 2022, pp. 3989–3999. [doi:10.1109/WACV51458.2022.00404](https://doi.org/10.1109/WACV51458.2022.00404)
- Kang, Le, Peng Ye, Yi Li, and David Doermann  
“Convolutional Neural Networks for No-Reference Image Quality Assessment.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740. [doi:10.1109/CVPR.2014.224](https://doi.org/10.1109/CVPR.2014.224)
- Mittal, Anish., Anush Krishna Moorthy, and Alan Conrad Bovik  
“No-reference image quality assessment in the spatial domain,” *IEEE Transactions on image processing*, 21 (12), 2012, pp. 4695–4708. [doi:10.1109/TIP.2012.2214050](https://doi.org/10.1109/TIP.2012.2214050).
- Su, Shaolin, Qingsen Yan, Yu Zhu, *et al.*  
“Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676. [doi:10.1109/CVPR42600.2020.00372](https://doi.org/10.1109/CVPR42600.2020.00372)
- Zhang, Weixia, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang  
“Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network,” *IEEE Transactions on Circuits and Systems for Video Technology* 30 (1), 2018, pp. 36–47. [doi:10.1109/TCSVT.2018.2886771](https://doi.org/10.1109/TCSVT.2018.2886771)



# Text and Layout Recognition for Tibetan Newspapers with Transkribus

Franz Xaver Erhard

(Leipzig University)

igitising 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.<sup>1</sup> The UK-German collaborative research project Divergent Discourses. Narrative Construction in Tibet, 1955–1962<sup>2</sup> takes advantage of these developments.

---

<sup>1</sup> 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.

<sup>2</sup> 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-20<sup>th</sup> 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 gсар '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 Hackett's work, Columbia University pioneered<sup>3</sup> this effort and, in cooperation with the Beinecke Rare Books and Manuscript Library at Yale<sup>4</sup>, the Musée Guimet, and the Collège de France, scanned their holdings of the *Tibet Mirror* and made them available freely between 2009 to 2013.<sup>5</sup> 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/>.

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

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

<sup>5</sup> A table of contents and images are available [https://openlibrary.org/works/OL17161360W/Yul\\_phyogs\\_so\\_so%CA%BEi\\_gсар\\_%CA%BEgyur\\_me\\_lon%CC%](https://openlibrary.org/works/OL17161360W/Yul_phyogs_so_so%CA%BEi_gсар_%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<sup>6</sup> alongside some annotated manual transcriptions.<sup>7</sup> 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 “ཏུ་མཚོན་ཀུན་རྩེས་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](https://www.tibetmirror.com/edition=key%3A/books/OL25732003M) (accessed September 15, 2024) or [https://archive.org/details/ldpd\\_6981643\\_000](https://archive.org/details/ldpd_6981643_000) (accessed September 15, 2024).

<sup>6</sup> 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/>

<sup>7</sup> See [https://salamandre.college-de-france.fr/archives-en-ligne/ead.html?id=FR075CDF\\_000IET002&c=FR075CDF\\_000IET002\\_de-310](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.<sup>8</sup> 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.<sup>9</sup> 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,<sup>10</sup> or later in the ornamental titles of many newspapers (Fig. 1), or as a decorative style for headlines and headings. However, the differences

---

<sup>8</sup> 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.

<sup>9</sup> 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).

<sup>10</sup> 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).

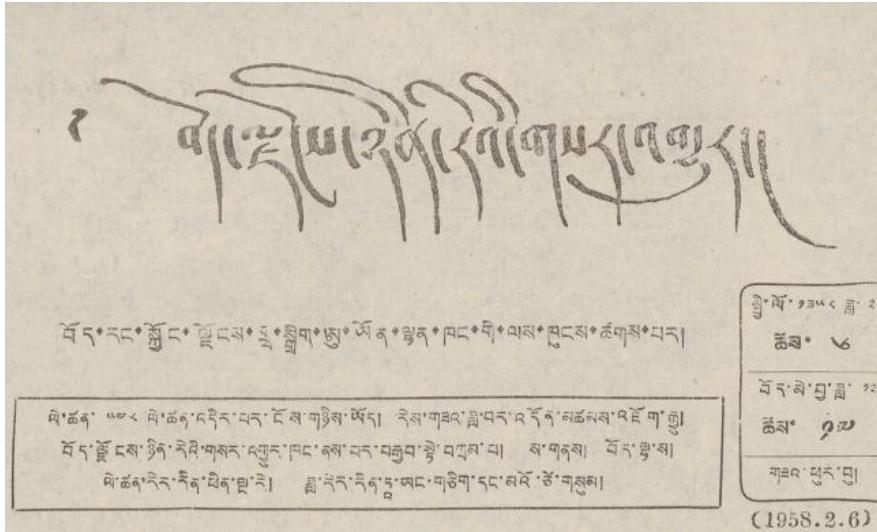


Figure 1 Masthead of the Tibet Daily (TID) February 6, 1958 with handwritten title by the 14th Dalai Lama (Hartley 2003: 87). (Oriental Institute, CAS, Prague XIV91-1959)

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 ལྷིལ་ལོ་ in standard *uchen* and ལྷིལ་ལོ་ in the *druitsa* (*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.<sup>11</sup>

<sup>11</sup> 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](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

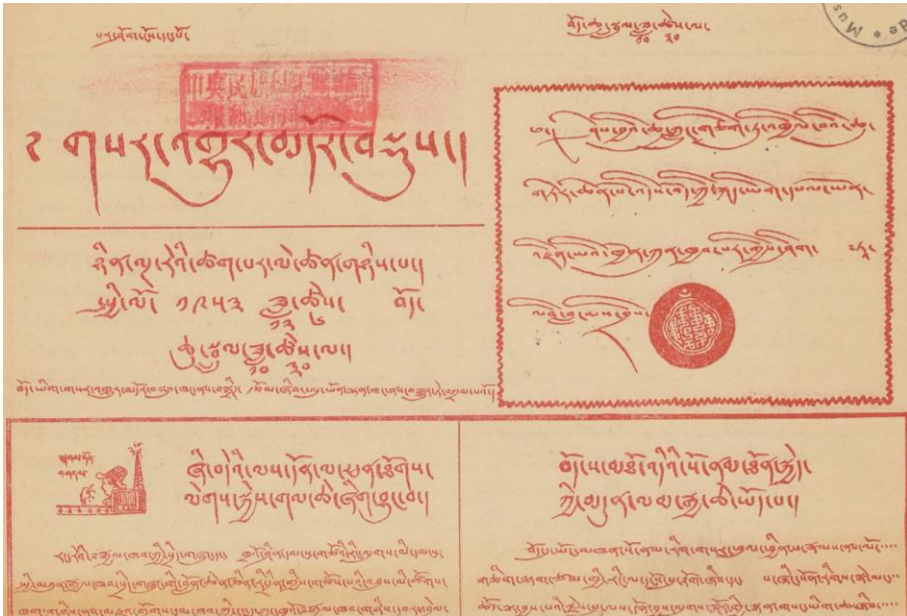


Figure 2 Example of different scripts in the News in Brief (NIB) from December 6, 1953 (Grassi Museum für Völkerkunde ASZAG9)

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 20<sup>th</sup> 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.<sup>12</sup>

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.<sup>13</sup> 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

---

<sup>12</sup> See Erhard & Hou 2018: 10. See also the discussion in Pistorius 2019: 11–12.

<sup>13</sup> 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,<sup>14</sup> 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 gсар 'gyur* ("Qinghai Tibetan Language News") starting in 1951<sup>15</sup> 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*).<sup>16</sup> 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

---

<sup>14</sup> No 25 of Appendix 1 in Sawerthal 2018: 345.

<sup>15</sup> No 22 of Appendix 1 in Sawerthal 2018: 345.

<sup>16</sup> 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,<sup>17</sup> while Transkribus offers its platform as a service. Therefore, Transkribus is more economical for small projects such as *Divergent Discourses*.<sup>18</sup>

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

---

<sup>17</sup> 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.

<sup>18</sup> 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.<sup>19</sup>

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:<sup>20</sup>

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

---

<sup>19</sup> 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.

<sup>20</sup> 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 Paṅḍita*, an 18th-century autobiography of a Tibetan aristocrat, collated from various manuscripts and published in two volumes in 1987 in Chengdu.<sup>21</sup> 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.

---

<sup>21</sup> Rdo ring 1987. For more on the text and its author see Erhard 2020a; Erhard 2020b.

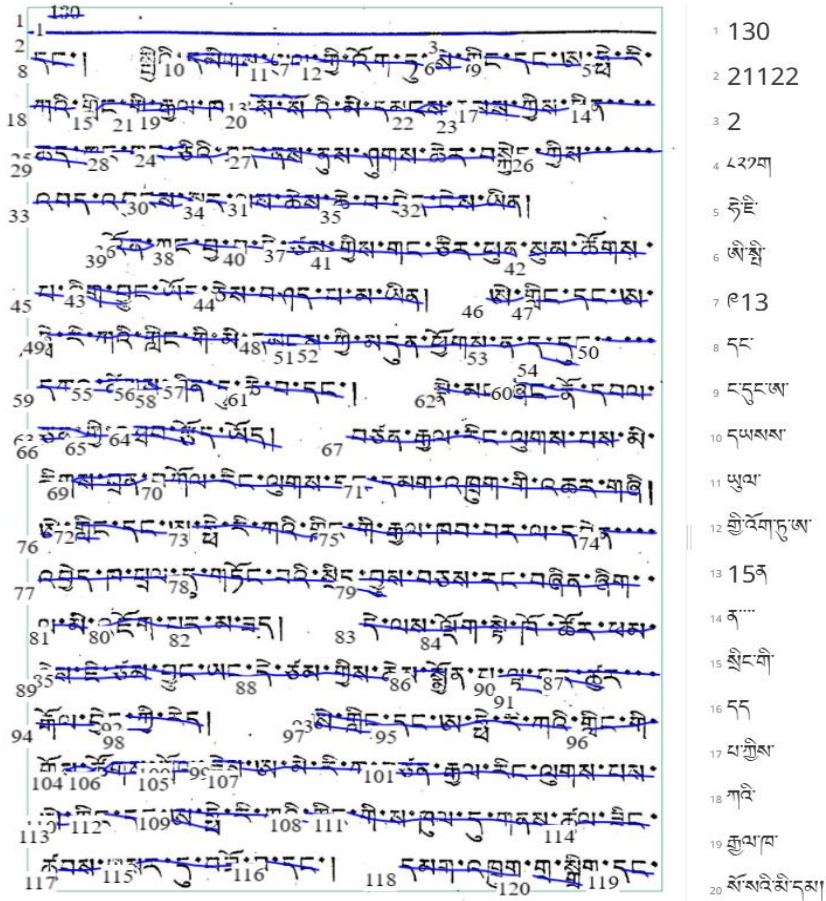


Figure 3 Results of HTR without previous baseline recognition

- (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 20<sup>th</sup> century in an iterative process starting with pages manually annotated during the creation of Ground Truth from excerpts from the *Biography of Doring Paṅḍita*.

Figure 4 Baseline default options in the Transkribus 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.<sup>22</sup>

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

---

<sup>22</sup> 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 ལྷོ་ལ་དགའ་བ་དང་། དཔའ་ཚལ་ཆེ་ཞིང་ཆོ་གྲོས་དང་།  
 2 ལྷན་པའི་མི་དམངས་ཡིན་པས་ལོ་རྒྱུད་ལྷག་ལ་གསུམ་ཀྱི་ཚོན་དུ་  
 3 རང་འདི་མེས་པོ་མེས་ཀྱིས་ལྷན་ལཱ་ལ་ལྷན་པའི་མིའི་རིགས་ལ་  
 4 ལྷན་པའི་ཆེ་ཞིང་འོད་སྔོན་ལ་ལྷན་པའི་རིག་གནས་གསར་  
 5 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་  
 6 ལྷན་པའི་ལོ་རྒྱུད་ (ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 7 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 8 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 9 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 10 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 11 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 12 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 13 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 14 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 15 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 16 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 17 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་  
 18 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་

- 1 ལྷོ་ལ་དགའ་བ་དང་། དཔའ་ཚལ་ཆེ་ཞིང་ཆོ་གྲོས་དང་།
- 2 ལྷན་པའི་མི་དམངས་ཡིན་པས་ལོ་རྒྱུད་ལྷག་ལ་གསུམ་ཀྱི་ཚོན་དུ་
- 3 རང་འདི་མེས་པོ་མེས་ཀྱིས་ལྷན་ལཱ་ལ་ལྷན་པའི་མིའི་རིགས་ལ་
- 4 ལྷན་པའི་ཆེ་ཞིང་འོད་སྔོན་ལ་ལྷན་པའི་རིག་གནས་གསར་
- 5 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་
- 6 ལྷན་པའི་ལོ་རྒྱུད་ (ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 7 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 8 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 9 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 10 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 11 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 12 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 13 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 14 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 15 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 16 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 17 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་
- 18 ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་པའི་ལོ་རྒྱུད་ལྷན་

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.

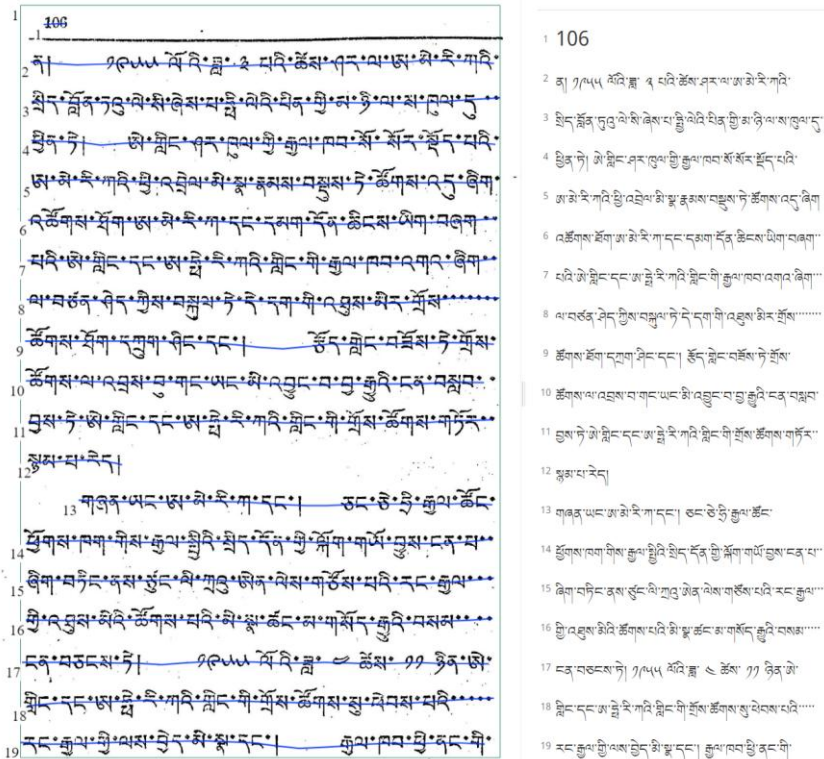


Figure 6 Transkribus screen shot showing perfect results for baseline detection and subsequent HTR with TMUP 0.1

### 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.<sup>23</sup>

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 Paṅḍita*, 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).<sup>24</sup> 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,

---

<sup>23</sup> 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.

<sup>24</sup> 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* ཨཔ་ (rendering the labial fricative “f”) in *hpha ran zi* ཨཔ་ར་མི (France), but also unusual orthography in Chinese names such as Le’u Hro’o chi ལེ་ལུ་ཧྲོ་འོ་ཅི (Liu Shaoqi 劉少奇), or the Chinese appellation *hru’u ci* ཧྲུ་ལུ་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.<sup>25</sup>

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.<sup>26</sup>

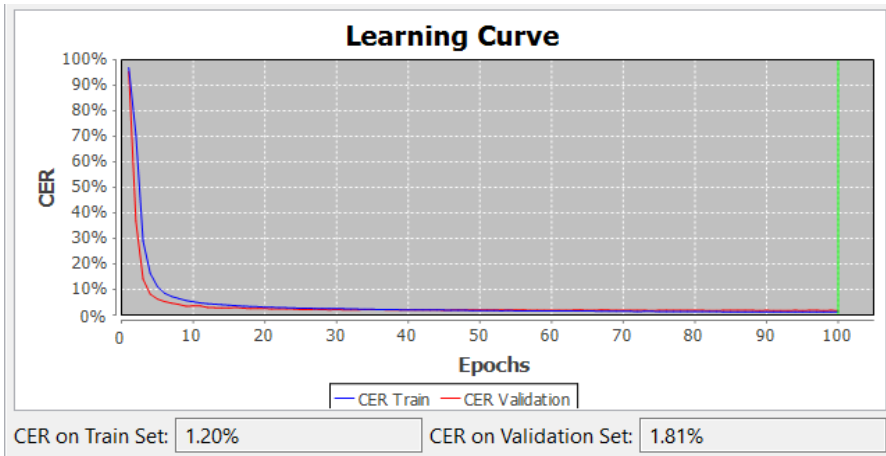


Figure 7: Tibetan Modern U-chen Print (TMUP) 0.1 Learning Curve

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.<sup>27</sup>

<sup>25</sup> 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.

<sup>26</sup> 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.

<sup>27</sup> 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་ཧི་ (for *cang ci hri* ཅང་ཅི་ཧི་) or *le ze thun ho zi gung zi* ལེ་ཟེ་ཐུན་ཧོ་ཟི་གུང་ཟི་ (for *we zi than ho zi gun zi* ངེ་ཟི་ཐུན་ཧོ་ཟི་གུང་ཟི་), occasionally confusing the very similar vowel signs for *i* ཧི་ and *e* ཅེ་ or also very similar syllables *nga* ཅང་ and *ja* ཅེ་. Another source of misreading stems from unclear print, such as *bcong* བཙོང་ (for *btson* བཙོང་ས) or *klog* ལྷོག (for *glog* ལྷོག).

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.<sup>28</sup> 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.

---

<sup>28</sup> 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.

(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<sup>29</sup>), initially showed a CER of 3.3%. The latest

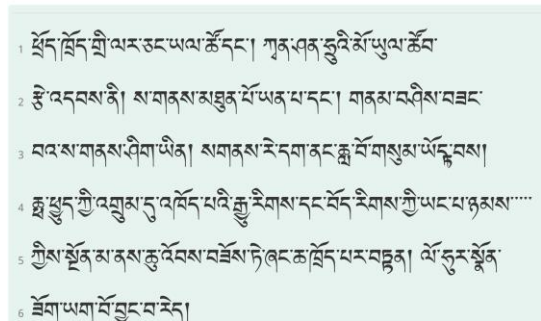
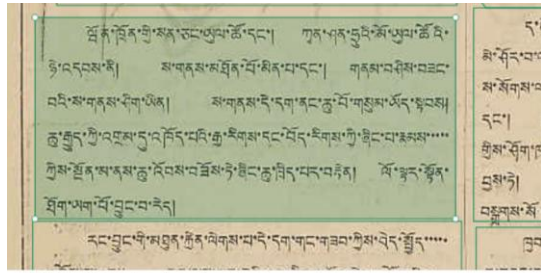


Figure 8 Performance of model minjiang v01 (ID 59357) with a CER of 0.5 on unseen text

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%.<sup>30</sup>

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

<sup>29</sup> 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.

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

<sup>31</sup> 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).<sup>32</sup> 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.<sup>33</sup> 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.<sup>34</sup>

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.<sup>35</sup>

---

<sup>32</sup> 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.

<sup>33</sup> For a comprehensive tag list, see the appendix, section 5.1.

<sup>34</sup> 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.

<sup>35</sup> 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).

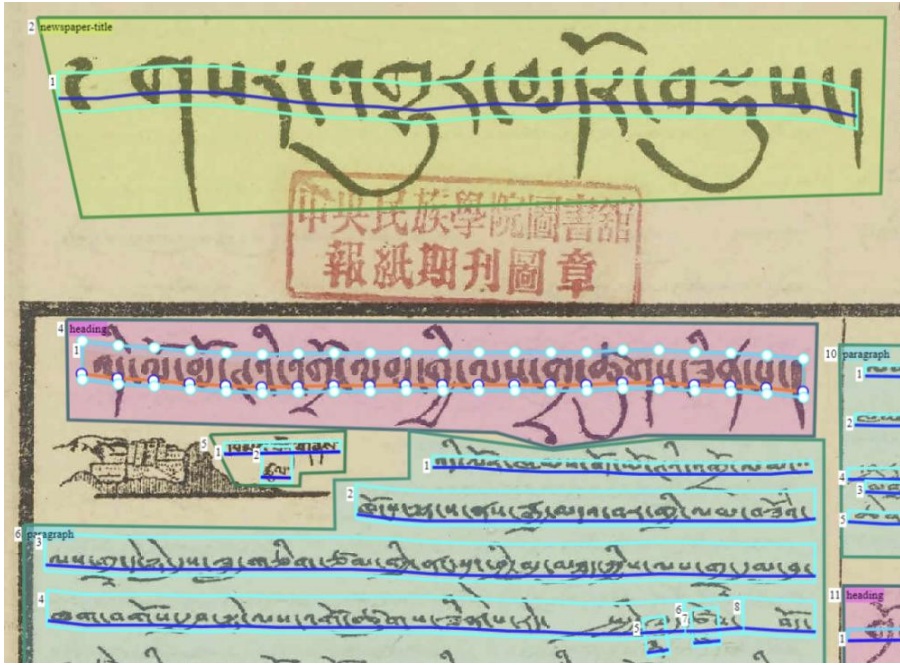


Figure 9 Line polygons (turquoise), automatically calculated from baselines (blue)

#### 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.<sup>36</sup>

<sup>36</sup> This approach was suggested by the Transkribus team following a roundtable on Transkribus for Asian Languages at TUC24 organised by Rachael Griffith and



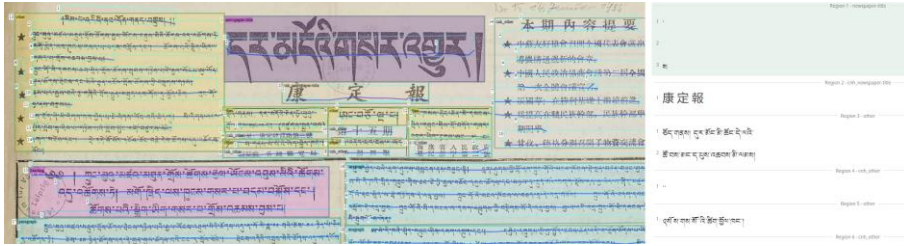


Figure 10 HTR after baseline recognition: Region 1, the newspaper-title, has been incorrectly recognised as three lines with narrow line polygons. Consequently, the HTR model has failed to transcribe this correctly.

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.

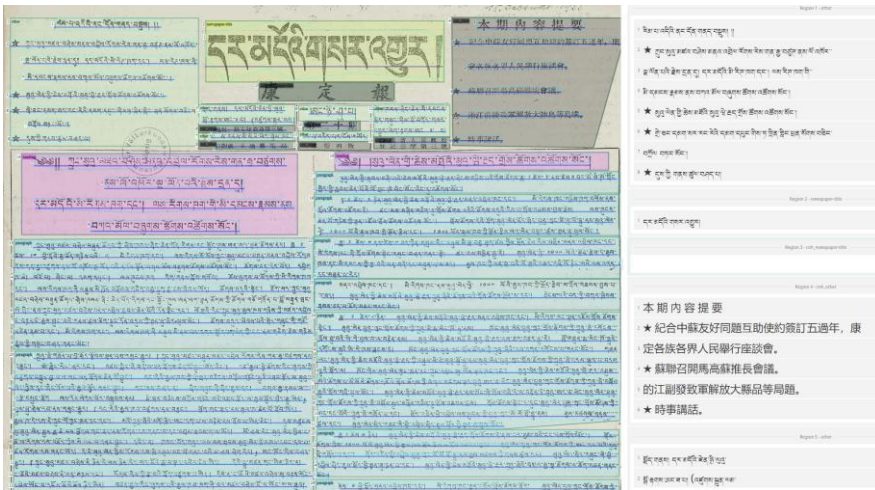


Figure 11 HTR after detection of line polygons with FM TibNewsLines 0.2.2:

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-corpora," *Zenodo*, 2022. [doi:10.5281/ZENODO.6513855](https://doi.org/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](https://doi.org/10.17875/gup2022-2054).

Bradburne, James M.

*Die schwarze Stadt an der Seidenstraße: Buddhistische Kunst aus Khara Khotu (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](https://doi.org/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) Asien*, 2024. [doi:10.48796/20240313-000](https://doi.org/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/digital-himalaya/collections/journals/ret/pdf/ret\\_72\\_03.pdf](https://d1i1jdw69xsqx0.cloudfront.net/digital-himalaya/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](https://doi.org/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](https://doi.org/10.5281/zenodo.10815272).

Li'u hra'o chis 劉少奇  
*Krung gor mar khe si dang le nyin ring lugs rnam 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](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](https://doi.org/10.1108/JD-09-2023-0183).

PaganTibet

"Reconstructing the Tibetan Pagan Religion," 2023. Available online <https://www.crao.fr/recherche/pagantibet-documenter-la-premiere-reconstruction-de-pratiques-preboudhiques-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](https://bdrc.org/W1PD96348).

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](https://doi.org/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](https://doi.org/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” ཅ for abbreviating the final consonants -gs འགས་ 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)*

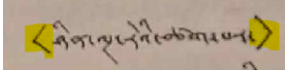
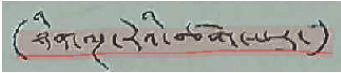
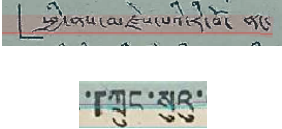
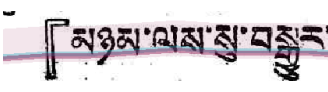
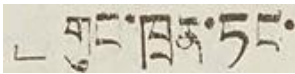
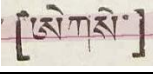
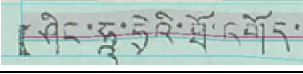
Abbreviations in other languages must be transcribed as in the original.

## 2 Transcribing parenthesis, bullet, and punctuation marks

## 2.1 Parentheseses 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



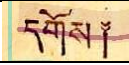



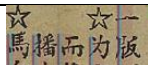
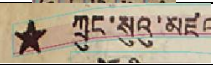
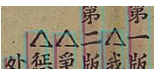


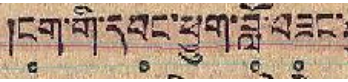
Description	Example	Transcription
angle brackets		< ... > U+003C / U+003E
round brackets		( ... ) U+0028 / U+0029
Chinese traditional single quotation mark		U+300C / U+300D
Chinese traditional double quotation mark		『...』 U+300E / U+300F
Chinese traditional vertical single quotation mark		┌...┐ U+FE41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

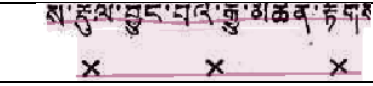
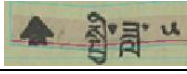
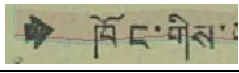
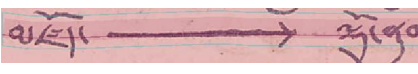
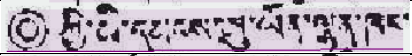

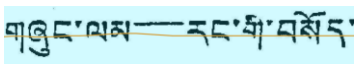
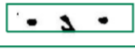
## 2.2 Bullets and punctuation marks

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
<i>yig mgo mdun ma</i> <i>yig mgo sgab ma</i>		☞ U+0F04 ☛ U+0F05
<i>che mgo</i> preceding the names of high incarnates		ལ U+0F38
<i>sbrul shad</i>		ལྷ U+0F08
<i>nyis tsheg shad</i>		། U+0F10
<i>rin chen spungs shad</i>		༎ U+0F11
<i>gter tsheg</i>		ེ U+0F14
<i>sgra gcan 'char rtags</i>		ེེ 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
<i>ngas bzungs sgor rtags</i> (Emphasis mark)		ེེ U+0F37

Name	Example	Unicode
<i>ku ru kha</i> (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*: ◁*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*).

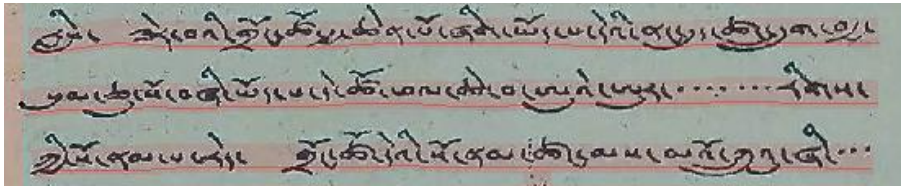


Figure 12 Intersyllabic tsheg and line/space-filling dots in *ume*

### 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

<p>“irregular” gap</p>		<p>SPACE</p>
<p>Short regular space</p>		<p>SPACE</p>

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)

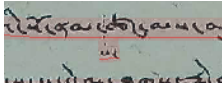
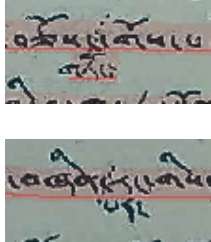
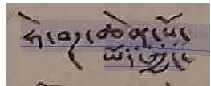
## 4 Corrections, additions, marginalia

## 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*

<b>Tag</b>	<b>Explanation</b>
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. <i>kham s yul ni</i>
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]

<b>Tag</b>	<b>Explanation</b>
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
CNH_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

<b>TAG</b>	<b>Explanation</b>
CHIN	Chinese text (left-to-right)
CHIN-VERT	vertical Chinese text
DBU-MED	Tibetan text in cursive script (incl. <i>dbu med</i> , <i>'khyugs</i> 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




# Foreign Names and Places in Tibetan Newspapers of the 1950s and 1960s

Franz Xaver Erhard (Leipzig University)

and

Xiaoying 笑影 (Leipzig University)

 Tibet was in close contact with neighbouring civilisations for all its known history. This cultural exchange tremendously impacted the religion, culture and language spoken on the Tibetan plateau. Buddhism was the most significant import, leaving a strong imprint in the Tibetan language, including many loanwords and Indic names. One of the more difficult things in the Tibetan language is naming foreign people and places beyond its semiotised cultural sphere. In past centuries, important places, such as along the Silk Road, in Mongolia, or India, received their own Tibetan names. Similarly, the Indian Buddhist deities, saints and sages received Tibetan names. However, in modern times and increasingly in the mid-twentieth century, the integration of Tibet into Communist China, Tibet and a Tibetan readership became exposed to an enormous number of Chinese and international place names and personal names.

In the Divergent Discourses project,<sup>1</sup> while preparing Ground Truth, i.e., accurate and verified transcriptions of samples of the project's newspaper corpus to train a text recognition model, we compiled a list of 1,299 foreign personal and place names found in the

---

<sup>1</sup> The project received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (<https://gepris.dfg.de/gepris/projekt/508232945?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/>.

newspapers alongside their original and Chinese forms.<sup>2</sup> We discovered, perhaps not too surprisingly, that many of these proper names were rendered in multiple ways into Tibetan, which is not only causing potential problems for our plan to digitally and automatically mine the newspaper corpus, but the variety of approaches to bring these names into the Tibetan language also tells us about the avenues these terms travelled and the practice of newspaper editing/translating as well as the state of minority language policies in the PRC of the 1950s and 1960s.

### 1 Introduction

Since ancient times, Tibet has been in close contact with and entertained cultural exchanges with its neighbours. This exchange doubtlessly has left its imprint in the form of loanwords and foreign place and personal names in the Tibetan language, or rather Tibetan languages, as various rules apply in the language variations of Amdo, Kham or Central Tibet. These loans had to be adapted to Tibetan phonetic and orthographic rules and eventually incorporated into the Tibetan lexicon. In this process, many of these foreign borrowings became increasingly difficult to distinguish from original Tibetan words over time. An oft-quoted example is དེབ་ཐེར་ *deb ther*, a Tibetan term for ‘annals’, which is a loan from the Persian دفتر *daftar*.

Borrowings from a wide range of languages are known, including Khotanese (Emmerick 1985), Mongolian, Turkish, Hindi, and English, but mostly alongside the import of Buddhism from Sanskrit and Chinese (Beyer 1993; Laufer 1919; Laufer 1916). As Bertold Laufer emphasised, many loanwords were introduced alongside Buddhism during the Tang dynasty, and their phonetic values represent contemporary Chinese pronunciation (Laufer 1916: 407).

---

<sup>2</sup> The dataset is divided into three separate csv-files containing (1) 96 Tibetan names, (2) 414 non-Tibetan foreign personal names, and (3) 787 non-Tibetan foreign place names, always including variants and the corresponding English or Chinese forms, see Erhard and Xiaoying 笑影 2024.

To represent foreign words, in particular loans from Sanskrit, and their pronunciation in the Tibetan language, several conventions emerged over time, such as the use of subscribed འ' to represent long vowels as in ཀཤལ་ *ka pā la* for Sanskrit *kapāla* or the five reversed letters (*log pa'i yi ge lnga*) to transcribe Sanskrit retroflexes as in *paṇḍita* པཎྌིཏྲ. A subscribed ཧ *h* represents aspirated voiced stops and affricates of the Sanskrit original.<sup>3</sup>

In modern language usage, interestingly, the fricative [f] in loanwords or foreign names is generally represented with the un-Tibetan ligature ཧ *h+pha*, as in ཧར་ཤི ཧ *h+pha ran zi* (France) or ཧེ་ལུ་ཤེ་ལེ་ ཧ *he lu'u sha'o h+phu* (Nikita Khrushchev).<sup>4</sup> The assumption that this is a recent invention of the PRC is supported by the fact that Goldstein (2001: 1176) lists only 29 entries starting with ཧ *h+pha*, all generally representing Chinese borrowings, while the Tibetan dictionary compiled by Lobsang Tendar (Blo bzang bstan dar 2010), published in India, lists no entry for ཧ *h+pha*.

Intriguingly, there is no uniform system for rendering foreign names in Tibetan. In the 18th-century *Biography of Doring Paṇḍita*, for example, the widely travelled author Tenzin Panjor (Bstan 'dzin dpal 'byor, 1760-c. 1811) gives vivid descriptions of the Chinese cities of Xining fu 西寧府, Lanzhou fu 蘭州府, Xi'an fu 西安府, or Chengdu fu 成都府. He names these places Zi ling hu ཟི་ལིང་ལུ, Khreng tu hu ཟེང་ཏུ་ལུ, etc, thus transcribing the character *fu* 府 with the Tibetan ལུ *hu*. At other places, the fricative [f] is rendered in Tibetan with the aspirated ཤ *pha* ཤམིར་ཀ *a phi ri ka* (Africa) (NIB 1955.06.04) or རེལ་ *ne pha* transcribing the acronym NEFA (North-East Frontier Agency) during the China – Indian war in Arunachal Pradesh (TIM 1963.02.01)

These preliminary remarks demonstrate that the Tibetan language has developed various techniques to include foreign terms and names in its vocabulary and showcase its ability to adapt to various cultural and linguistic influences over time. Borrowings then are also not stable

<sup>3</sup> Peter Schwiieger gives a comprehensive overview of letters used to render Sanskrit in Tibetan (2006: 23).

<sup>4</sup> This ligature is referred to in modern textbooks from the PRC as the "one additional letter" (*kha snon yi ge gcig*), see e.g. (Skad yig 1993: 22).

but undergo change influenced by social and political tidings. While in the first half of the 20<sup>th</sup> century, British and British-Indian culture and the English language had significant influence as many technical terms, such as རི་ལུ་ *ri li/lu* > rail, or མོ་ཏ་ *mo ta* > motor = car, amply demonstrate, in the second half of the 20<sup>th</sup> century, this influence is superseded by the influx of Chinese concepts and linguistic borrowings. Hence, the now obsolete English borrowing ཕེ་སེ་ཀོབ་ *pe se kob* > bioscope, cinema,<sup>5</sup> which was replaced by གློག་བརྟན་ *glog brnyan*, explained by Beyer (1993: 138) as a loan translation of the modern Chinese 电影 *dian ying*.

## 2 *The list of place names and personal names: selection criteria and sources*

The entries in this list were randomly selected while transcribing sample pages from the newspapers. It goes without saying that this list is far from being complete or representative of all the choices of the newspaper editors of the period.

We anticipated complications with unsystematic orthography in the representation of the names of countries, regions, cities and other places as well as personal names of authors, politicians and statesmen in our newspaper corpus. We created the list (and are still adding names) to allow for normalisation or enhance Named Entity Recognition (NER) at a later stage. Therefore, we collected non-Tibetan place names and personal names as well as names of persons with both a Tibetan and a Chinese name, such as སངས་རྒྱས་ཡེ་ཤེས་ *Sangs rgyas ye shes* 天寶 Tian Bao or Tempa Landoo བསྟན་པ་ལུན་འགུབ་ *Bstan pa lhun 'grub* 丹巴隆舟 alias 高攀桂 Gao Qianguì.

The list of foreign person and place names was extracted from the following newspapers:

---

<sup>5</sup> Nowadays generally out of use, the term ཕེ་སེ་ཀོབ་ *pe se kob* has only made it into one dictionary, Goldstein and Narkyid (1986: 69), which lists it as a borrowing from Hindi and synonym to *glog brnyan*.

## (1) People's Republic of China:

- *Mtsho sngon bod yig gsar 'gyur* Qinghai Tibetan News (QTN),
- *Kan lho'i gsar 'gyur* South Gansu News (SGN),
- *Dar mdo'i gsar 'gyur* Kangding News (KDN),
- *Dkar mdzes nyin re'i gsar 'gyur* Ganze Daily News (GDN),
- *Ming kyāng tshags dpar* Minjiang News (MJN),
- *Bod ljongs nyin re'i gsar 'gyur* Tibet Daily (TID),
- *Gsar 'gyur mdor bsdus* News in Brief (NIB)

## (2) India:

- *Yul phyog so so'i gsar 'gyur me long* Tibet Mirror (TIM)
- *Rang dbang gsar shog* Freedom (FRD),
- *Rang dbang srung skyob gsar shog* Defend Tibet's Freedom (DTF),
- *Krung dbyang gsar 'gyur* Central Weekly News (CWN),
- *Bod mi'i rang dbang* Tibetan Freedom (TIF).

The name list contains 1,299 entries in total. The larger part are place names with 787 entries with their spelling variations. "Indonesia" appears here with 13 different spellings. Among the 417 personal names, Zhou Enlai is the entry with the most variants.

### 3 *Transcription and its variants in Tibetan newspapers of the 1950s and 1960s*

The newspapers in the Divergent Discourses project's corpus stem from a period of tremendous political and social change. The newspapers report on local events and the several provinces of the PRC and beyond in the world. Besides new political concepts of democracy, communism, feudalism, etc., new technologies such as "combine harvester" (KDN 1955.05.28), "aerial seeding" (QTN 1959.01.28), "veterinary disease control" (QTN 1959.03.04), a plethora of geographical names, political and administrative titles and names of politicians, presidents and others are for the first time transcribed



into Tibetan in these papers. Despite the tendency for centralised control, the newspapers' authors, editors, and translators utilise the approaches outlined above to the best of their ability to render all these new terminologies, place and personal names in Tibetan. However, they do not follow a unified system, and various approaches become evident when flipping through the newspapers' pages. Additionally, there are multiple renditions of some names, sometimes even in the same publication. In the extreme case, we found as many as 12 different ways to render the name of the important Communist leader Zhou Enlai in Tibetan.

### 3.1 Transcription of Chinese names

For the same Chinese name, several different transliterations are found. The example of the Chinese Communist politician and long-term prime minister Zhou Enlai 周恩來 (1898–1976) demonstrates that the transcription into Tibetan was not following a unified system but was done by individual authors or translators:

- ལུ་ཨེན་ལེས་ *kru'u en les* (Note: final *sa* is not an agentive marker)
- ལུ་ཨེན་ལེ *kru'u en le*
- ལེ་ཨེན་ལེ *kre'u en le*
- ལེ་ཨེན་ལེན་ *kre'u en lan*
- ལེ་ཨེན་ལེན་ *kra'u an lan*
- ལེ་ཨོ་ཨེན་ལེ *kre'o en le*

Additionally, as a comparison of newspapers of different regions shows, the authors or, more likely, translators based their transcriptions on the Tibetan target dialect of their publication. For example, we can differentiate transcriptions based on Amdo pronunciation in the *Mtsho sngon bod yig gsar 'gyur* (QTN): ལེ་ཨེན་ལེ *krig nin le*. This writing reflects the Amdo dialectal pronunciation of what, again, is the local Chinese dialect of Xining (tʂu ən̄ lɛ) rather than the Beijing dialect-based Mandarin (tʂou ən̄ lai).

The *Min kyāng tshag dpar* (MJN) offers transcriptions based on Kham pronunciation again reflecting the Barkham local Chinese dialect pronunciation of Zhou Enlai (tsəu ɲən nai):

- གྲིབ་གླིང་ལཱི་ *kri'u gin le'i*
- གྲིབ་གླིང་ལཱི་ *kri'u gin la'i*

The transcriptions found in the *Tibet Mirror* (TIM) published in Kalimpong, on the other hand, are based on Hindi or, ultimately, English pronunciation:

- ཅའོ་ཨེན་ལཱི་ *ca'o en la'i*
- ཅའོ་ཨེན་ལཱི་ *ca'o an le*
- ཅའོ་ཨེན་ལཱི་ *ca'u en la'i*

### 3.2 Transliteration of non-Chinese names

When dealing with non-Chinese names, a variety of different approaches are widespread. For example, the Russian Khrushchev (Хрущёв, 赫魯曉夫 *he lu xiao fu*), the name of the First Secretary of the Communist Party of the Soviet Union from 1953 to 1964 Nikita Khrushchev (1894-1971), is transcribed in three different ways:

First, editors attempted to phonetically transfer the Chinese 赫魯曉夫 *he lu xiao fu* resulting in the variations already described above:

- ཧེ་ལུ་ཤའོ་ཕུ་ *he lu'u sha'o h+phu*
- ཧེ་ལུ་ཤའོ་ཕུས་ *he lu'u sha'o h+phus*
- ཧེ་ལུ་ཤའོ་ཕུས་ *he la'u sha'o h+phus*
- ཧེ་ལོ་ཤོ་ཕུ་ཕུ་ *he lo'u sho h+phu'u*
- ཤེ་ལུ་སྐོ་ཕུ་ *she lu'u skyo h+phu*
- ཧེ་ལུ་ཤའོ་ཕུ་ *he lu sha'o phu*
- ཧེ་རུ་ཤའོ་ཕུ་ *he ru sha'o phu*

Second, transcriptions attempt to imitate the Russian pronunciation, but struggle with the consonant clusters: ཕུ་ཤེ་ཅོབ་ (*khu she cob*) or ཕུ་ཤེ་ཅོབ་ (*khu*

*shi cob*) seek to phonetically transcribe “khru-sh-chev” dividing up the consonant ལ “shch” across two Tibetan syllables. Similarly, in the Amdo dialect-based transcription, the Russian consonant cluster xp “khr” is split into two syllables in ཀེརུདྲི (*ke ru dri*), rendering “kh-ru-shchev”. In both cases, additional vowels are inserted in deference to Tibetan phonotactics.

Similarly, the toponym Việt Nam (Vietnam 越南 *yue nan*) is rendered in different ways, sometimes following various Chinese dialects, as demonstrated by the following examples:

- ཡོནན *yo nan*
- ཡོའོནན *ya’o nan*
- ཡུལན *yu lan*
- ཡུནན *yu nan*
- གཡོདནའན *g.yod na’an*
- གཡོདནླན *g.yod nān*
- གཡོདནན *g.yod nan*

Alternatively, transcriptions imitate the Vietnamese or English pronunciation:

- རློའ་ནམ *wi ta nam*
- རློའ་ནམ *wet nam*
- རིདིནམ *bi di nam*
- རིད་ནམ *bi da nam*
- རིནམ *bi nam*

Regarding personal names in the wider Sinosphere, most Japanese and Koreans officially write their names in Chinese characters. Each Chinese character has a distinct Chinese, Japanese, Korean or Vietnamese pronunciation. In Tibetan newspapers, such names are generally transcribed following the Chinese characters. The Japanese Kishi Nobosuke 岸信介 would be pronounced as An Xinjie in Chinese. The Tibetan རནཞིནཀེ *nan zhin ke* (QTN) derives from the Chinese pronunciation An Xinjie. Another example is the name of Kim Il-sung 김일성 (1912–1994), the “eternal president” of North Korea, has the

Chinese form 金日成 *jin ri cheng*. The Tibetan transcriptions ཅིན་རི་ཅེང་ *cin ri khreng* or ཅིན་རི་འཛིན་ *cin ri drin* (NIB) follow the Chinese.

Chinese name forms were discarded in Vietnam at the end of the 19th century. Still, the practice was followed by a few persons, mostly in cultural contexts. So, for example, the Vietnamese Politician Lê Thanh Nghi (1911–1989) is quoted in Tibetan newspapers as ལིས་ཅིང་ཡིང་ *lis ching ying* (GDN), a Tibetan transcription of his Chinese name 黎清毅 *li qing yi* (GDN).

### 3.3 Multiple Origins for Transcriptions for the Same Place

For many countries, several Tibetan names are circulating, which have different origins. Thailand, for example, is frequently transcribed following the English toponym “Thailand” with several slight variations:

- ཐཱ་འི་ལང་ *tha'i lanḍ*
- ཐཱ་འི་ལན་ཏེ *tha'i lanḍ*
- ཐཱ་ལན་ཏི *tha lan ḍi*
- ཐཱ་ལན་ཏ *tha lan ḍa*
- ཐཱ་ལེན *the len*
- ཐཱ་འི་ལན *tha'i lan*

At the same time Thailand is found in Tibetan newspapers denoted by its pre-1939 name Siam, rendered in Tibetan as སི་ཡམ་ (*si yam*) or ལྷ་ཡུལ་ (*shyam yul*). The latter name is a mixed form and combines the phonetic transcription ལྷ་ *shyam* with ཡུལ་ *yul*, a Tibetan term for “country”. This approach is also followed in the mixed borrowing from Chinese 泰國<sup>6</sup> *tai guo* where alongside the full transfer ཐཱ་ཀོ་ *the ko*, or ཐཱ་ཀོ་ *the go*, the

---

<sup>6</sup> The traditional character 國 is used in our newspapers before 1959, and the simplified form 国 after 1959.

transfer མེའི *the'i* (Thai) is combined with རྒྱལ་ཁབ་ *rgyal khab*, the Tibetan gloss for country to the mixed form མེའི་རྒྱལ་ཁབ་ *the'i rgyal khab* (Thailand).<sup>7</sup>

Sri Lanka is well known in classical Tibetan literature under the name ལང་ཀ་ *lang ka* or ལང་ཀ་ལྷ་ *lang kā* or Simhala (སིང་གལ་ *sing ga la*) which is still used in Tibetan newspapers, e.g. the Central Weekly News (CWN), together with the variants སིང་ལའི་གླིང་ *sing+gal 'i gling*, སང་གི་ལ་ *sanggil* (TIM), and སེང་གལ་ *seng+ga la*. Nevertheless, in other instances, the newspaper editors opted for a modern transcription of Sri Lanka based on the Chinese 斯里蘭卡 *si li lan ka*: སི་རི་ན་གླ་ *si ri na gha*.<sup>8</sup> Finally, there are also transcriptions of the older English name Ceylon as སེ་ལོང་ as well as of the Chinese borrowing 錫蘭 *xi lan* as ཤེ་ལང་ *shi lang*, ཤེ་ལན་ *she lang*, or ཤེ་སེ་ལོང་ *shis lon*.

The Vietnamese capital Hà Nội is phonetically transcribed from Vietnamese as ཧོ་ནོ་ *he no* or from its Chinese name form 河内 *henei* as ཧོ་ནོ་འེ་ *ho na'e* or ཧོ་ནོ་འི་ *ho ni*. Alternatively, we also found the old name of the city Đông Kinh (東京 *dong jing*, literally 'eastern capital') phonetically transcribed following its Vietnamese pronunciation as ཧོང་ཁུན་ *stong khun*.

An interesting case is "Germany" most often transcribed phonetically from English as ཇར་མན་ *jar man* or འཇར་མན་ *'jar man*, or from the Chinese 德國 *de guo* as ཇོ་གོ་ *de go* or ཇོ་གོ་ *te go*. In the newspapers of the 1950s and 1960s, we found the now mostly obsolete forms ཇོ་ཡི་སྐྱི་ *de yis kri*, ཇོ་ཡི་སྐྱི་ *ti yi kri*, or ཇོ་ཡི་སྐྱི་ *te yis kri* probably deriving from the German 'deutsch' ('dɔɪ̯tʃ) via Chinese 德意志 *de yi zhi*.

### 3.4 Coexistence of Transcription and Translation

The previous section dealt with phonetic transcriptions from different etymologies, all imitations of respective pronunciations. This section

<sup>7</sup> Alternatively, the syllable *the'i* could be read as *the* (Thai) plus possessive '*i*', changing the phrase to "country of the Thai [people]". According to Stephan Beyer the combination of a "transfer with a native gloss on the meaning of the transferred element" (1993: 145) is well attested in examples such as in *rma bya* (peacock) combining Skrt. *māyura* and Tib. *bya*.

<sup>8</sup> In some Chinese dialects, especially in Sichuan and Yunnan, 蘭 *lan* is pronounced as *nā*, *næ*, or *nan*. It is also not uncommon to see mix-ups in the pronunciation of *l* and *r*. For example, Krang Dbyi sun (2000) gives སི་ལི་ལན་ཀ་ *si li lan kha* for Sri Lanka.

showcases the differing approach to linguistic borrowing that focuses on the meaning and attempts translations in full or part of the original term.

- The county Maoxian 茂縣 phonetically transcribed in Tibetan newspapers as མོང་ཤན *mong shan*, མོང་ཤེན *mong shen*, མོང་ཞན *mong zhan*, མའོ་ཤི་འན *ma'o shi'an*, but also partly translated as in མའུ་རུང་མོང་ *ma'u rdzong*.
- Hainan Island 海南島 *hai nan dao* is transcribed as ཧེ་ནན་ཏོ་འོ་ *he nan to'o*, but also partly translated as in ཧེ་འི་ནན་གླིང་ཕྱར་ *he'i nan gling phran*.
- Europe 歐洲 *ou zhou* is transcribed as ཡུ་རོབ་གླིང་ *yu rob gling*, or ཡོ་རོབ་གླིང་ *yo rob gling*, ཡའོ་རུབ་གླིང་ *ya'o rub gling* as well as phonetically གི་ལུ་གྲི་ལུ་ *gi'u kri'u*, རུང་ལུ་ལུ་ *rgu'u kru'u*, or རུང་ལུ་ལུ་ *rgu'u khru'i*.

The examples above show the coexistence of phonetic transcriptions alongside partial translations. Particularly, the components “county” 縣 *xian*, “island” 島 *dao*, or “continent” 洲 *zhou*, which usually form a part of the toponym in Chinese (e.g. “Maoxian” cannot simply be referred to as “Mao”), were translated into Tibetan as རུང་མོང་ *rdzong*, གླིང་ཕྱར་ *gling phran*, and གླིང་ *gling* respectively.

An example of a full translation of a place name is the Inner Mongolian capital of Hohhot (ᠬᠣᠬᠣᠲᠤ) rendered 呼和浩特 *hu he hao te* in Chinese and subsequently phonetically transcribed into Tibetan as ཧུ་ཧོ་ཧོ་འོ་ *hu ho ha'o the* or referred to in its translation from Mongolian as མཁར་སྒོན་མོ་ *mkhar sngon mo*, the Blue City.

### 3.5 Different Names in Tibetan and Chinese

Almost all regions with Tibetan populations have places with two or more unrelated names. Most of these places are located in frontier regions with high ethnic diversity. Culturally or politically dominating ethnic groups maintained their own place names independent of other ethnic groups' names. Additionally, many frontier places received Chinese names during the Qing Dynasty. In 1706, the county ལྷག་མ་ཐམ་ཁ་ *lcags zam kha* in today's Ganze TAP, Sichuan, was named 瀘定 *lu ding* by Emperor 康熙 Kangxi (r. 1661–1722). In the newspapers, however,

besides the Tibetan name also the phonetic transcription of the acquired Chinese name ལུ་ཏིན *lu'u tin* is used.

Some examples, to name but few, are the famous trade hub of Kangding 康定 in Sichuan, which is known as དར་རྩེ་མདོ་ *dar rtse mdo* in Tibetan. Further to the north is Luhuo 爐霍, otherwise known as བླ་མགོ་ *brag mgo*. In Qinghai, རེབ་གོང་ *reb gong* (occasionally རེབ་གོར་ *reb kong*), the home of the Tibetan monk-scholar Gendün Chöphel (*dge 'dun chos 'phel* 1903–1951), is known in Chinese as 同仁 *tong ren* and often referred to in Tibetan newspapers in the phonetic transcription ཐུན་རིན་ *thun rin* or ཐོར་རིན་ *thong rin*. In today's Tibetan Autonomous Region, Yadong 亞東, the trade port on the route to India, is known as གོ་མོ་ *gro mo* in Tibetan. In Yunnan, today's Shangri-La (香格里拉 *xianggelila*, སེམས་ཀྱི་ཉི་མོ་ *sems kyi nyi zla*) until 2001 used to be known in Chinese as Zhongdian 中甸 and རྒྱལ་ཐང་ *rgyal thang* in Tibetan.

### 3.6 Confusions

Historical influences, dialectal varieties and homophony in both source and target language provided a broad spectrum of linguistic borrowings and approaches to transcription. At the same time, they also were the source of confusion and mistakes: The names of the Chinese provinces 陝西 Shaanxi and 山西 Shanxi, are homophones in Chinese with only the tone of the first character differing. Consequently, they are transcribed identically in Tibetan (མཁན་ཤེས་ or མཁན་ཤེས་) but refer to different locations depending on the newspaper. Only the context can clarify whether 陝西 Shaanxi or 山西 Shanxi is referred to.

In general, in newspapers such as the News in Brief (NIB), the Minjiang News (MJN), or the Central Weekly News (CWN), the Chinese province Yunnan 雲南 is referred to in the newspapers as ཡུན་ནན་ *yun nan*, ཡུན་ནན་ *yun na'an*, ཡུལ་ནན་ *yul na'an*, ཡུན་ནན་ *yun nān*, or ཡུན་ལེན་ *yun len* based on the Chinese name of the province. In the *Tibet Mirror* (TIM), however, the Tibetan term ཡུན་ནན་ *yun nan* is used to refer to Greece in the form ཡུན་ནན་རྒྱལ་ཁབ་ *yun nan rgyal khab*. While the general transcription of Greece is based on the Chinese 希臘 ཧེ་ལ་ *he la* ཧེ་ལ་ན་ *he lan* or English གི་རི་སི

*gi ri si*, the choice of ཡུན་ནན་ *yun nan* (TIM) seems to be derived from the Hindi यूनान (*yūnān*) or Persian یونان (*yūnān*) term for Greece.

#### 4 Conclusion

On first sight, the implications of these varying transcriptions seem mostly relevant to questions of normalisation (see article Kyogoku *et al.* 2025 in this issue), which is a necessary process in the creation of a digital corpus that ensures that, e.g., all searches return a full list of results, but more importantly, that all processes run on the corpus yield reliable results. However, our preliminary and hand-compiled list of just over a thousand foreign place and personal names and their Tibetan transcriptions also allows us to draw a few conclusions about the socio-historical context of the transcriptions and the newspapers from which they were culled.

First, although for centuries engaged in cultural and political exchange, Tibet was not prepared to easily and smoothly digest the massive influx of Chinese terms in the 1950s and 1960s, many of which also represented new political or social concepts.

Second, the coinage of new borrowings in the form of translations or transcriptions was left to individual translators, resulting in an enormous variety of transcriptions based on (a) the source language or dialect and (b) the target Tibetan dialect corresponding with the newspaper's distribution area.

Third, this individualisation of transcription practices highlights two points: For one, translations and transcriptions into Tibetan initially did not follow a centralised system (as is well established today with committees and publications that control the coinage of new terms). It also shows how long and difficult it is for such a system to emerge and solidify. But more importantly, it suggests that the developments in the 1950s and 1960s were extremely rushed and did not leave much time for planning and standardisation. To quickly bring out the message its translation was left to local translators at the cost of a unified vocabulary. In our understanding, this underlines the



importance of mass communication, in our case, newspapers, in the formative years of the PRC in its Inner Asian frontier regions.

### Bibliography

Beyer, Stephan V.

*The classical Tibetan language*. Bibliotheca Indo-Buddhica series 116. Delhi: Sri Satguru, 1993.

Blo bzang bstan dar (ed.)

*Bod kyi tshig mdzod chen mo*. [Great Tibetan Dictionary]. Dharamsala: Bod gzhung Shes rig las khungs, 2010.

Emmerick, Ronald Eric

“Tibetan loanwords in Khotanese and Khotanese loanwords in Tibetan.” In G. Gnoli and L. Lanciotti (eds.) *Orientalia Iosephi Tucci Memoriae Dicata*. 3 vols, pp. 301–317. Serie Orientale Roma. Roma: Istituto italiano per il Medio ed Estremo Oriente, 1985.

Erhard, Franz Xaver and Xiaoying 笑影

“Toponyms and Anthroponyms from Tibetan-language Newspapers of the 1950s and 1960s. Three Name Lists,” *Zenodo*, 2024. [doi:10.5281/zenodo.14526124](https://doi.org/10.5281/zenodo.14526124)

Goldstein, Melvyn C. and Ngawangthondup Narkyid (eds.)

*English-Tibetan Dictionary of Modern Tibetan*. Reprint of the edition Berkeley 1984. Dharamsala: Library of Tibetan Works and Archives, 1986.

Kyogoku, Yuki, Franz Xaver Erhard, James Engels, and Robert Barnett

“Leveraging Large Language Models in Low-resourced Language NLP: A spaCy Implementation for Modern Tibetan,” *Revue d’Etudes Tibétaines* (73), 2025, pp. 187–220.

Krang Dbyi sun 張怡蓀 (ed.)

*Bod rgya tshig mdzod chen mo*. [Great Tibetan Chinese Dictionary]. Pe cin: Mi rigs dpe skrun khang, 2000.

Laufer, Berthold

"Loan-Words in Tibetan," *T'oung Pao* 17 (1) 1916, pp. 403–542.

*Sino-Iranica: Chinese Contributions to the History of Civilization in Ancient Iran*. Field Museum of Natural History: Anthropological Series XV, No. 3. Chicago: Field Museum, 1919.

Schwieger, Peter

*Handbuch zur Grammatik der klassischen tibetischen Schriftsprache*. Beiträge zur Zentralasienforschung 11. Halle (Saale): International Institute for Tibetan and Buddhist Studies, 2006.

*Skad yig: deb dang po*. [(Tibetan) Language. First Book]. Lo dgu'i 'gan babs slob gso nyin hril po'i lam lugs kyi bod rang skyong ljongs slob chung slob deb. Lha sa: Bod ljongs mi dmangs dpe skrun khang, 1993.



# Leveraging Large Language Models in Low-resourced Language NLP: A spaCy Implementation for Modern Tibetan

Yuki Kyogoku (Leipzig University), Franz Xaver Erhard (Leipzig University), James Engels (University of Edinburgh) and Robert Barnett (SOAS University of London)

**L**arge Language Models (LLMs) are transforming the possibilities for developing Natural Language Processing (NLP) tools for low-resource languages. While languages like Modern Tibetan have historically faced significant challenges in computational linguistics due to limited digital resources and annotated datasets, LLMs offer a promising solution. This paper describes how we leveraged Google’s Gemini Pro 1.5 to generate training data for developing a basic spaCy language model for Modern Tibetan, focusing particularly on Part-of-Speech (POS) tagging. Combining traditional rule-based approaches with LLM-assisted data annotation, we demonstrate a novel methodology for creating NLP tools for languages with limited computational resources. Our findings contribute to the broader effort to enhance digital accessibility for low-resource languages while offering practical insights for similar projects in computational linguistics.

## 1 Introduction

Despite recent advancements in digital humanities, low-resource languages, such as Tibetan, still face substantial challenges due to limited digital resources and tools. Addressing these gaps is essential

for enhancing digital accessibility and supporting advanced linguistic research.

A key step in Tibetan language processing is the development of computational tools such as POS taggers, which underpin many NLP applications. However, creating such tools for Tibetan poses particular difficulties, including a lack of large-scale datasets, inconsistencies in existing text sources, and the distinct syntactic features of the language.

To address the scarcity of Tibetan-language corpora, a diverse dataset was compiled from contemporary sources, and we took additional steps to normalise text and correct inconsistencies such as punctuation errors and the use of abbreviations. Furthermore, integrating external tools, such as the Botok tokeniser (see section 3.1 below), is explored to manage structural challenges typical for the Tibetan language, particularly the absence of spaces between Tibetan words.

Finally, we demonstrate how using Google's LLM Gemini Pro 1.5 for automated data annotation critically contributed to developing a Tibetan language model for spaCy. Moreover, we showcased the potential of LLMs to contribute to the development of NLP resources more generally, aligning with the broader objective of addressing the digital divide for low-resource languages.

## 2 *The Scope of the Research*

The Divergent Discourses project<sup>1</sup> investigates the construction of narratives in Tibet in the 1950s and 1960s. It is interested in extracting

---

<sup>1</sup> The Divergent Discourses project received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (<https://gepris.dfg.de/gepris/projekt/508232945?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/>. We also would like to thank Michael Richter and Tyler Neill for their valuable feedback on an earlier draft of this paper.

such narratives and related elementary information — such as agents, places, events, and their relationships — from digitised Tibetan newspapers. Although Automatic Text Recognition (ATR) is a critical component of our broader project (see Erhard 2025 in this issue), this paper focuses specifically on the process of training a basic spaCy language model for Modern Tibetan, using the digitised texts generated through ATR alongside digital-born materials.

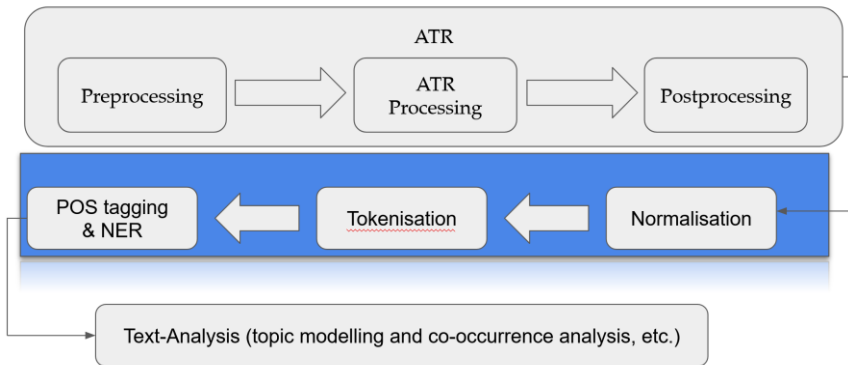


Figure 1 The complete project pipeline, with the scope addressed in this paper highlighted in blue. It is important to note that this figure should not be confused with the training process of a POS tagger.

The open-source natural language processing library spaCy is written in Python and Cython. It enables the training of pipelines for tasks such as POS-tagging, parsing, and Named Entity Recognition (NER) from scratch, even for languages not natively supported.<sup>2</sup> We chose to work with spaCy because it is the underlying NLP engine for a wide range of corpus mining and analysis tools, such as Wordless,<sup>3</sup> or the integrated Leipzig Corpus Miner (iLCM),<sup>4</sup> a tool offering various analytical functionalities such as topic modelling, co-occurrence analysis, and supervised text classification. The training process for the

<sup>2</sup> See <https://spacy.io/usage/spacy-101> (accessed January 25, 2025).

<sup>3</sup> Wordless is an integrated corpus tool with multilingual support for studying language, literature, and translation developed by Ye Lei (叶磊), Shanghai. See <https://github.com/BLKSerene/Wordless> (accessed January 25, 2025).

<sup>4</sup> <https://ilcm.informatik.uni-leipzig.de/> (accessed December 18, 2024), see also Niekler 2014, 2023.

spaCy model comprises two main phases: the generation of segmented and annotated training data and the subsequent training of the spaCy language model.

The structure of this study is organised as follows: First, the availability of NLP tools, such as tokenisers, POS taggers and related technologies, is discussed, with a focus on tools existing for the Tibetan language. This discussion also includes an overview of the authors' contributions, explicitly developing a Tibetan text normaliser and adapting the Botok tokeniser for the modern Tibetan language. Secondly, we examine the annotated corpora of Tibetan training data in general and emphasise the need to create a new Modern Tibetan corpus. Subsequently, we describe how an LLM such as Google's Gemini Pro 1.5 can be leveraged for the automatic creation of training data and thus overcoming the main obstacle of low-resourced languages: the lack of training data. The final section details the process of training a Modern Tibetan language model for spaCy.

### 3 *Relevant Tools for Tibetan*

This section discusses the techniques and digital tools relevant to the objectives of *Divergent Discourses* and examines existing digital tools for Tibetan. Despite recent advances in NLP, certain limitations persist in Tibetan digital humanities. These limitations are examined in detail below.

#### 3.1 *Tokenisation*

Tokenisation is the process of breaking down a larger text into smaller units, called tokens, which can, depending on the context, be words, subwords, sentences, or even individual characters. However, in most cases, tokens correspond to what are commonly referred to as "words." Tokenisation is a fundamental step in NLP and computational linguistics and the first step in preparing text data for analysis or model training. Its quality directly impacts the accuracy of

subsequent steps, such as POS-tagging or NER. Thus, developing, selecting or customising an appropriate segmentation tool, known as a tokeniser, is a critical step in Tibetan language processing.

Unlike languages such as English, which use white spaces to separate words, Tibetan — often described incorrectly as monosyllabic — presents challenges for NLP because of its writing system and the absence of word delimiters. Therefore, a tokeniser is needed to divide strings of syllables (morphemes) into tokens.<sup>5</sup>

One notable example of a Tibetan tokeniser is Botok, developed by OpenPecha.<sup>6</sup> Botok is designed as a rule-based tokeniser and was initially created to process Classical Tibetan texts. However, it is customisable, allowing users to adapt its functionality to specific text genres by altering its dictionary file. This adaptability makes Botok valuable for working with Tibetan texts in various contexts. The details of these customisations will be described in section 4.2 below.

### 3.2 POS-Tagging

In computational linguistics, a language model often needs to identify the grammatical roles of words in a text, such as nouns, adjectives, and other POS. This process, known as POS-tagging, is a critical step for many NLP tasks that require an understanding of syntactic distributions and permissible word combinations based on the grammatical features of a language. POS-tagging enhances the accuracy and efficiency of complex NLP tasks such as text parsing, machine translation, and text generation. In the context of Divergent Discourses, POS-tagging is essential because it facilitates the development of NER, which identifies words representing entities such as places, dates, people, or organisations. The pre-identification of proper nouns through POS-tagging simplifies the NER process.

---

<sup>5</sup> For those who are interested in tokenisers, see [https://huggingface.co/docs/transformers/v4.29.1/tokenizer\\_summary](https://huggingface.co/docs/transformers/v4.29.1/tokenizer_summary) (accessed December 18, 2024).

<sup>6</sup> <https://github.com/OpenPecha/Botok> and [https://github.com/Esukhia/botok-data/tree/master/dialect\\_packs](https://github.com/Esukhia/botok-data/tree/master/dialect_packs) (accessed December 18, 2024).

While POS-tagging is not mandatory for all NLP tasks — topic modelling (see Schwartz & Barnett 2025 in this issue) and semantic search (see Engels & Barnett 2025 in this issue), for example, often rely on word embeddings instead, which do not need to have POS-tags — many widely-used NLP platforms, including spaCy and Stanford CoreNLP,<sup>7</sup> require POS-tagging as part of the engine for newly trained language models, regardless of their specific architecture.

Tagging, however, is a computationally complex task. While some words consistently represent a single part of speech, others are homographic (e.g., “the sailor closed the hatches” vs. “the hen hatches an egg”) or polysemous across syntactic categories (e.g., “I hurt my back” vs. “the prime minister said he would back the new policy”). Addressing these challenges typically involves one of two approaches: (1) a rule-based system, which relies on predefined linguistic rules, or (2) a machine-learning-based system, which uses pre-annotated data to infer patterns. In the following subsections, we discuss various types of POS-taggers and introduce those specifically designed for Tibetan.

### 3.2.1 *The Rule-based Approach*

The rule-based approach to tagging, used by most early POS-taggers for any language (including the earliest taggers in English), seeks to replicate a human’s immediate linguistic intuitions by applying complex tree-based rulesets in which either specific morphologies or orthographic environments dictate a certain classification.

The earliest Tibetan taggers (and many taggers still in use by the few Tibetan NLP researchers and users) were rule-based, using computational versions of older syntactic descriptions of Tibetan. This is the case with Hackett (2000), which applies Wilson’s (1998) general rules of Tibetan syntax. This tagger operates essentially like a decision tree: locate the word > check the word in the dictionary > decide which

---

<sup>7</sup> <https://stanfordnlp.github.io/> (accessed July 12, 2024).



POS is most likely given the immediately preceding or neighbouring POS > assign the relevant tag.<sup>8</sup>

The Buddhist Digital Resource Centre (BDRC) and Esukhia developed a more complex form of rule-based tagging using a modified rule-based approach. Botok, in addition to tokenisation, also applies POS tags to its tokens by looking for those tokens in an internal dictionary and returning the most likely result based on the morphological features of the word and its neighbours. The following diagram describes the general architecture of Botok's tagger:

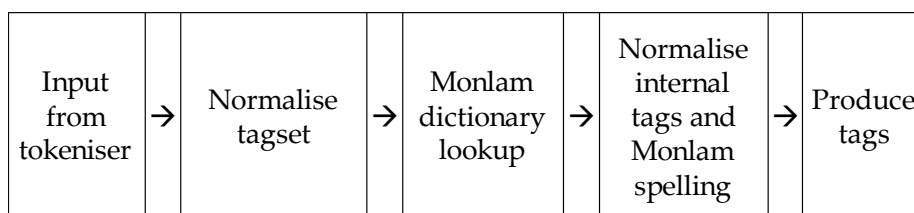


Figure 2 General architecture of Botok's tagger.

This tagger, however, does not allow to any significant extent for conditional probabilities based on the syntactic environment of the token it tries to tag. Most of the time, the tag defaults to the first listed POS in Monlam's online dictionary.<sup>9</sup> To measure the overall quality of Botok's POS tagger, we compared its POS tags on a small sample text to those of an expert consultant who is both a Tibetologist and a theoretical linguist (see Appendix A).<sup>10</sup> We compared each tag to the expert's gold-standard tag set, using a sliding window of two tokens to accommodate splitting decisions involving syllables where even experts disagree. In a 30-word sample paragraph, compared to the expert reviewer, Botok's POS-tagger performed to a middling degree,

<sup>8</sup> Detailed information about the internal function of the tagger is not included in the original publication, and the source code is not easily retrievable online.

<sup>9</sup> <https://monlamdictionary.com/> (accessed December 18, 2024).

<sup>10</sup> We thank Dr Camille Simon (Paris) for her assistance with this process.

with an agreement accuracy of 53.3% (16/30).<sup>11</sup> Botok's tagging facility is thus relatively elementary compared to its tokenising capability.<sup>12</sup>

Other rule-based tagging systems exist for Tibetan. A system for syntactical analysis called Constraint Grammar, for example, has been successfully built into Tibetan taggers (see Faggionato & Garrett 2019, Garrett & Hill 2015), yielding better results than a simple dictionary-lookup system such as the one included with Botok. However, systems based on Constraint Grammar lack the learning flexibility that one would find in a machine-learning model.

Rule-based taggers are, in general, limited by a few critical weaknesses independent of resource availability:

- (1) Unless combined with adaptive conditional probabilities, rule-based taggers are entirely context-independent. Thus, when a tagger of this type encounters an unfamiliar syntactical structure, it is liable to make an incorrect decision or to apply a tag such as "other," "unknown," or "none."
- (2) A subset of the context-independence problem is the problem of polysemy: how should the parser decide to tag a word that can, at different times, function as a noun, verb, adjective, or something more fine-grained such as a past participle?
- (3) Unstructured text, or text in unfamiliar genres, can confuse rule-based taggers that are overfitted to certain textual environments or contexts. For example, suppose the rules on which a tagger is

---

<sup>11</sup> This score was achieved with the aid of a sliding window (since Botok tends to be "splitty", i.e., it tokenises longer compounds in its elementary syllables) and after allowing for Botok's consistent quasi-errors (almost everything Botok considers "other" is a noun, for instance). In addition, Botok's tagset is more limited than the Universal Dependencies (UD) tagset we instructed our annotator to use, and even for humans, it is arguable how to assign POS-tags in the UD tagset to such Tibetan linguistic objects as the genitive case ending, and we therefore give Botok some leeway, such as when it assigned PART to tokens our annotator tagged as AUX or ADP.

<sup>12</sup> It should be noted that Botok was developed with tokenisation in mind, and its POS-tagging capabilities are more of a side effect, according to one of the developers (Personal conversation with Hélios Drupchen Hildt, Diverge project meeting SOAS London, 26 August 2024).

based have been developed based on the conventions found in literary or religious texts. In that case, it will have difficulty tagging newspaper text. Social media, for example, a prominent use-case for NLP applications, poses a significant challenge to rule-based taggers because of the unending linguistic creativity of the public in informal linguistic contexts.

- (4) If dialectal variation is a concern, such as if the target language is pluricentric (Hindi-Urdu, Serbo-Croatian), and particularly when that variation consists of minor differences in morphology and syntax, these differences must be explicitly compensated for in the ruleset.

Prior to the development of LLMs and other transformer-based systems, rule-based taggers were generally the only tools available for Tibetan, despite its millions of speakers and a relative abundance of online resources. In addition, the amount of readily available annotated training data in Tibetan is unusually low compared to other languages of its size and reach.

### 3.2.2 *Machine-learning Approaches to Tagging*

There has nevertheless been an evolution in POS-tagging for Tibetan from a rule-based approach to an approach based on “shallow” machine learning. This was demonstrated in the case of Tibetan by the ACTib tagger developed by Meelen and Hill (2017), which combines a rule-based with a memory-based tagger.

ACTib performs word and sentence segmentation and provides POS-tagging, offering a high level of detail. It is designed explicitly for Classical Tibetan. The combination of a rule-based approach with the Tilburg Memory-Based Tagger (van der Sloot & van Gompel 2024) allows for more flexibility and yields better results for a wide range of texts than a simple rule-based system (Meelen 2021). ACTib has trained the Tilburg Memory-Based Tagger for use with Classical Tibetan, but it could be retrained for use with modern Tibetan. However, due to a lack of appropriate training data, it is impossible to

retrain the memory-based tagger of ACTib for our specific needs. Consequently, ACTib and its output did not meet the requirements of our study.

### 3.2.3 *Transformer Approaches to Tagging*

The shallow machine-learning approach may now become outdated by the major improvements represented by the development of transformers and transformer-based approaches to NLP tasks.<sup>13</sup> Tibetan LLM technology, such as the projects underway at Monlam AI,<sup>14</sup> are rapidly improving and will sooner or later gain an “intuitive” understanding of Tibetan.<sup>15</sup> GPT-4 is already capable of intuitively POS-tagging text from medium-resource languages like Vietnamese with nearly 100% accuracy. By September 2024, it could POS-tag a Tibetan sentence with around 80-90% accuracy. A comparison of POS-tagging in Tibetan by GPT-4, Claude 3.5 Sonnet, and Gemini shows (besides several allowable disagreements, mostly resulting from variations in their approach to tokenising and from variations in their tag sets) a similar error rate of 10-20%, which is low given the likely rate of progress (see Appendix F).

However, all LLMs are black boxes, and their output is difficult to predict. They hallucinate, occasionally add irrelevant information and generally share a tendency to be inconsistent, i.e., sporadically making differing tagging or translation decisions. We found that if the exact

---

<sup>13</sup> A transformer is a type of deep learning model widely used in tasks like language translation, text generation, and summarisation. Unlike earlier models, which process text sequentially (word by word) and apply limited contextual awareness to decision-making, transformers use an attention mechanism that allows them to focus on the most relevant parts of a sentence or sequence. This helps them understand the relationships between words more effectively, capturing context and meaning even across long distances within the text.

<sup>14</sup> <https://monlam.ai/about> (accessed December 18, 2024), and <https://github.com/MonlamAI> (accessed December 18, 2024).

<sup>15</sup> T-LLaMA, a first fine-tuned LLM for Tibetan based on META's LLaMA2, was published in December 2024, but could not yet be tested by Divergent Discourses, see Lv *et al.* 2025.

text is repeatedly submitted for tokenisation, the results are not always identical, as would be expected. Consequently, to leverage LLMs' intuitive knowledge of Tibetan, an equally powerful mechanism needs to be installed that can control and minimise inconsistencies as well as the rare yet unavoidable hallucinations so characteristic of LLMs.

Several dedicated POS taggers have been developed for Tibetan using Deep Learning and Transformer-based approaches. Li *et al.* (2022) combine deep learning techniques, specifically Bidirectional Long Short-Term Memory (Bi-LSTM) and Iterated Dilated Convolutional Neural Network (IDCNN), with machine learning methods, such as Conditional Random Fields (CRF), to propose an end-to-end model for joint Tibetan word segmentation and POS-tagging. Similarly, Xiangxiu *et al.* (2022) employ Embeddings from Language Models (ELMo) and the self-attention mechanism of Transformers to address challenges related to polysemy and out-of-vocabulary words. However, while the results of these Transformer-based POS-taggers are promising, none of the tools is publicly available, preventing their practical use.<sup>16</sup>

### 3.3 Normalisation of Texts

Normalisation is the process of reducing the randomness of a text such as various encodings, unnecessary characters to a set standard. This can also include more complex tasks such as stemming stripping, i.e., eliminating affixes, or lemmatisation, i.e., reducing variants to a base form. Tibetan, unlike languages such as Russian or Sanskrit, is an agglutinative language. This allows for direct searches of specific words, or at least of their stems, except in the case of conjugated verbs. However, several factors can complicate direct search functionality. For example, suppose traditional Tibetan brackets or numerals are found in the text alongside their non-Tibetan counterparts. In that case,

---

<sup>16</sup> BERT models have been developed by Tibet University and made available on Hugging Face (<https://huggingface.co/UTibetNLP>, accessed December 18, 2024). However, they are not specifically designed for POS-tagging.

these will be missed by a search for the standard, non-Tibetan version. The *tsheg* (།), when — as is common practice — used repeatedly as a filler at line breaks, may break up a search string and make a search unfeasible. Abbreviations, prevalent in early 20th-century Tibetan publications such as the *Tibet Mirror*, can also confuse a standard search engine. Only with the replacement of lead typesetting in the late 20th century with computerised typesetting did abbreviations become less frequent. To address these issues, the project developed a normaliser that converts such features into modern, standardised forms (Kyogoku *et al.* 2024a). It replaces traditional brackets and numerals with standard modern ones, e.g., the traditional Chinese quotation mark (「...」) is replaced by (“...”). It resolves simple abbreviations, such as the common contraction of final -ལས by -ལ, but also resolves complex abbreviations, such as འདྲེན་ into འདྲེན་གསུམ་ by referring to a list extracted from our newspaper corpus and a revised version of the list of more than 6,500 abbreviations compiled by Bruno Laine.<sup>17</sup> The normaliser also corrects errors such as improper punctuation. For instance, if a *tsheg* is missing between the letter *nga* (ཨ) and the punctuation mark *shad* (།), the normaliser automatically inserts a *tsheg* to ensure text consistency. These normalisations do not alter the raw or original text in the corpus but are saved in a new version that allows a user to retrieve a more complete set of results from a single search.

#### 4 Corpus of Training Data for Modern Tibetan

A text corpus of training data is a digitised collection of language data with specific linguistic annotation designed for computational analysis. The CoNLL-U format is often employed in linguistic research among the various formatting standards for such corpora. This format, commonly used in Universal Dependencies (UD), presents annotated sentences with POS tags and dependency labels, etc., in the CoNLL-U

---

<sup>17</sup> <http://www.rkts.org/abb/index.php> (accessed December 18, 2024).

format.<sup>18</sup> UD sets of POS tags have been developed for many languages, including classical ones. In addition, platforms like spaCy require their training data to be formatted in CoNLL-U. As a result, the CoNLL-U format has become the standard for syntactic and morphological annotation. We, therefore, formatted our training data in the CoNLL-U format to ensure compatibility for subsequent training of our spaCy Tibetan language model.

As highlighted in the introduction, Tibetan, like many low-resource languages, so far has limited amounts of data available for use in training. In addition, a UD set has not yet been developed for either Modern or Classical Tibetan. The requirements and structure for a Classical Tibetan treebank were outlined by Faggionato and Meelen (2019), but such a resource – again – would have limited applicability to our mid-20th-century Tibetan texts. Although some CoNLL-U-formatted files exist for Classical (Faggionato *et al.* 2021) and Modern Tibetan (Dakpa *et al.* 2021), their quality and volume are insufficient for training an effective POS tagger.

#### 4.1 Automatic generation of a training dataset for modern Tibetan

For the reasons outlined in the previous section, we decided to create our own training dataset so that we would then be able to train a spaCy model for modern Tibetan.

The initial step in creating a CoNLL-U file with Tibetan POS tags was the collection of a raw text corpus sufficiently large enough to train a spaCy language model. Practical constraints limited the corpus size to 100–200 MB. Data sources included Ground Truth transcriptions of Tibetan newspapers from the 1950s and 1960s, book publications from the same period, contemporary Tibetan newspapers from South Asia, and scraped content from openly accessible Tibetan news websites, particularly *Tibet Daily* (Kyogoku *et al.* 2024b).

Prior to the widespread use of LLMs, training datasets had to be manually annotated by human annotators. However, by mid-2024

---

<sup>18</sup> <https://universaldependencies.org/format.html> (accessed December 18, 2024).

LLMs such as ChatGPT, Gemini, or, more recently, Claude 3.5 Sonnet (released June 20, 2024), although not specifically trained on Tibetan material, were showing varying yet significant degrees of accuracy in their uses of Tibetan language. Since using LLMs as annotators for low-resource languages can save considerable time and effort compared to traditional manual methods, we opted to use LLMs to generate the necessary annotations automatically. We evaluated three LLMs, including ChatGPT, and identified Gemini Pro 1.5 (released February 15, 2024)<sup>19</sup> for its – at the time – superior accuracy and accessibility in Google's cloud environment as the best option for POS tagging Tibetan sentences (Appendix F and Barnett & Engels 2025: 25-28, 36-38 in this issue). To generate an annotated dataset appropriate for spaCy training, we implemented several optimisations in the prompt, i.e., the instruction or text provided to a LLM to elicit a specific response:<sup>20</sup>

- (1) **Entry selection:** We limited the automatic annotation to essential fields in the prompt: ID, form, and UPOS (Universal Part-of-Speech Tag), as a complete CoNLL-U format was unnecessary for our objectives.
- (2) **English translation:** Recent studies, such as Huang (2023), have demonstrated that incorporating English translations of non-English languages, particularly those classified as low-resourced, can significantly enhance performance in tasks related to language understanding, reasoning, and generation. Building on this finding, alongside Tibetan text, requests for English translations are included in the prompt to provide additional context, in order to improve the accuracy of the linguistic annotations.
- (3) **Pre-segmentation of sentences:** We found inconsistencies in Gemini's word segmentation of modern Tibetan texts. We

---

<sup>19</sup> <https://ai.google.dev/gemini-api/docs/models/gemini> (accessed December 18, 2024) and <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#performance> (accessed December 18, 2024).

<sup>20</sup> There are other useful techniques for improving the LLM performance. See Chen et al. (2023) and Deshpande et al. (2024).



therefore used Botok to segment our data prior to inputting it into the model, ensuring that token boundaries were preserved as needed. For this study, a customised “dialect pack” of Botok was used (see the detailed explanation in section 4.2. below).

- (4) **Structured prompt specification:** We specified a CoNLL-U-compatible format within the prompt, with entries separated by tabs, and extracted only the tab-separated line. This ensured that the output contained solely the table without any additional notes occasionally appended by Gemini. For instance, despite the differences in the output in Figures 3 and 4, only the tab-separated entries, i.e., the entries in the table (token number, token, English translation, and POS-tag) were extracted and saved in CoNLL-U format; additional content was disregarded.

**Response**

1	གངས་འབྲུགས་	snow_and_ice	NOUN
2	ལུས་རྩལ་	sports	NOUN
3	པས་	by	ADP
4	འགྲན་རར་	competition	NOUN
5	ཟོམ་ཚོག་པའི་	can_show	VERB
6	གྲུབ་འབྲས་	achievement	NOUN
7	གཏོད་པ་དང་	achieved	VERB

Figure 3 A response generated by Gemini Pro 1.5, presented in a table format without any additional annotations.

- (5) **Limiting the Prompt Size:** An increase in the length of the prompt can cause interruptions in Gemini’s answer generation. To address this, we had to ensure a reasonable prompt length and segmented the input Tibetan sentence using the *shad* (།) as a delimiter (note: this delimiter does not always signify the end of a sentence; see Figure 5).

**CoNLL-U format for the Tibetan passage:**

Token Number	Token	English Translation	POS-Tag
1	གངས་འབྲུག་གས་	glacier	NOUN
2	ལུས་རྒྱལ་པ་	athletes	NOUN
3	ས་	by	ADP
4	འབྲུག་རྒྱལ་	competition	NOUN
5	འདི་	in	ADP
6	ལྟོ་སྟོན་པ་	show, demonstrate	VERB
7	འདི་	that can be	PART
8	ལྷན་འབྲེལ་	results	NOUN
9	མཐོང་པ་	achieved	VERB
10	དང་	and	CCONJ

**Notes:**

- Some POS tags might be debatable depending on the context and interpretation.
- The English translations are approximate and provided for understanding, not as a one-to-one mapping.
- The CoNLL-U format typically includes more columns (lemma, morphological features, etc.), but only the requested ones are provided here.

Figure 4 Response with additional annotations generated by Gemini Pro 1.5

(6) **Correction of Annotation Inconsistencies (Post-Processing):**

The results produced several inconsistencies, such as the genitive suffix or particle, which was variably labelled as ADP or PART.<sup>21</sup> We addressed these discrepancies by applying standardisation rules in the post-processing stage, using a Python script. Additionally, if a single token in a sentence contained an unexpected POS-tag, not conforming to the UD standard, the whole sentence was eliminated by the script from the dataset. Consequently, the volume of data Gemini generates

<sup>21</sup> As for the POS-tags, see <https://universaldependencies.org/u/pos/index.html> (accessed January 15, 2025).

is smaller than the initial input dataset, albeit with a higher POS-tag accuracy.<sup>22</sup>

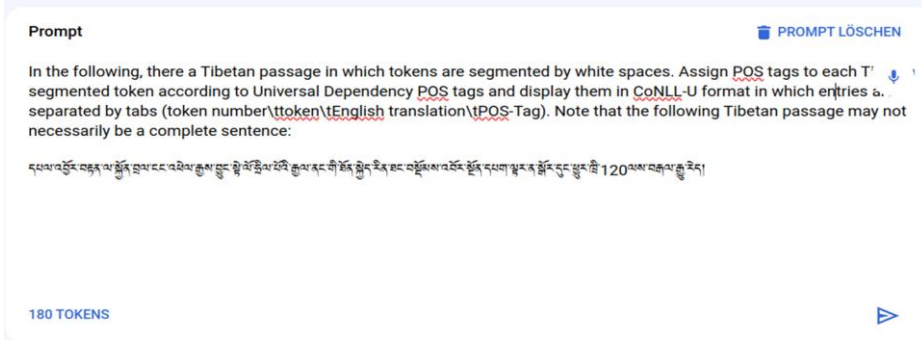


Figure 5 Example of a prompt. Note that this prompt was manually created in a browser, while our approach involves automatically generating prompts.

We found that in general, one-shot or few-shot prompts — where one or a few examples of the desired output are provided in the prompt — tends to perform better than zero-shot prompting, which does not include examples (Sivarajkumar 2023). However, this approach was not feasible because of the limitations of the prompt’s length (see point 5 in the paragraph above). Still, the evaluation of the training dataset produced by Gemini Pro 1.5 yielded promising results. A sample of 100 sentences was randomly extracted and evaluated by an expert in our team: 93% of the tokenisations and 91% of the POS-tags were deemed to be accurate (see Appendix B).

<sup>22</sup> The spaCy tutorial page (<https://spacy.io/usage/training>, accessed January 15, 2025) states, “If you want to train a model from scratch, you usually need at least a few hundred examples for both training and evaluation” The generated dataset (Kyogoku *et al.* 2024b) comprises approximately 50 MB of data (training data 39.3 MB + validation data 9.6 MB), and the dataset contains over 80,000 sentences (training data 64927 sentences + validation data 16231 sentences). The average number of tokens per sentence is approximately 7.5 (training data 7.41 + validation data 7.25 tokens).

#### 4.2 Customisation of Botok

When we first reviewed the results, we noticed that Gemini's tokenisation decisions, while generally acceptable, lacked consistency in some cases. However, consistent data is essential for training a spaCy language model. Consequently, as outlined in the third listed optimisation, we opted to enforce consistent tokenisation on Gemini by providing it with text pre-tokenised with Botok.

To enable support for Modern Tibetan, we used Botok's functionality, which allows users to adapt it to more specific language varieties by including "dialect-packs," which incorporate purpose-specific dictionaries. We created a "dialect-pack" for Modern Tibetan by modifying Botok's dictionary file (*tsikchen.tsv*).<sup>23</sup> We merged seven Modern Tibetan dictionaries with the original dictionary provided by Botok.<sup>24</sup> Duplicate entries were removed during this process.

However, the initial version of the resulting dictionary contained numerous compounds consisting of multiple words. Since our approach prioritises tokenising words into their smallest possible units, we removed compounds formed by genitive particles or those where adjectives or numbers modified nouns. Additionally, we limited the maximum number of syllables per word to four, as most Tibetan words fall within this range. This customised dictionary enabled tokenisation that aligns more effectively with our research objectives (Erhard *et al.* 2024).<sup>25</sup>

---

<sup>23</sup> As for how to customise dialect packs, see <https://github.com/OpenPecha/Botok?tab=readme-ov-file> (accessed January 15, 2025).

<sup>24</sup> The custom dictionary was compiled from Christian Steinert's collection (<https://github.com/christiansteinert/tibetan-dictionary/tree/master/input/dictionaries/public>, accessed January 12, 2024) and contains the following dictionaries: Grand Monlam Dictionary (default dictionary of Botok), [Jim Valby](#), [Ives Waldo](#), [Dan Martin](#), [Tshig mdzod chen mo](#), [Dung dkar](#), and [Tibetan Terminology Project](#). The resulting dictionary was cleaned up and edited to the project's requirements by removing double entries, phraseologisms, ungrammatical entries, etc. Moreover, ca. 1000 personal and place names compiled from the project's material were added; see Erhard and Xiaoying 2025 in this issue.

<sup>25</sup> The evaluation of the customised Botok, referred to as Modern Botok, is presented in Table 1 in Section 5.

### 4.3 Challenges in the generation of the training dataset

Some significant challenges remain unresolved in the process of generating a training dataset using Gemini Pro 1.5. One issue was the incorrect recognition of proper nouns foreign to the Tibetan language, as illustrated in Figure 6. This problem is mainly limited to proper nouns introduced to Tibetan through Chinese, particularly those which imitate Chinese pronunciation.<sup>26</sup> Most of these names are originally Chinese and consist of three syllables, as in the example of ལི་ཀེ་ཁྱང་ (Li Keqiang 李克强), the former Chinese premier (2013–2023). A potential solution involves incorporating proper names into the dictionary file of the Modern Botok dialect-pack (tsikchen.tsv); however, this has not yet been fully implemented.<sup>27</sup>

```
## sent_id = 1
# text = ལི་ཀེ་ཁྱང་ ལྷན་ རྒྱུ་ རྒྱུ་ ལྷན་ ལྷན་ ལྷན་ ལྷན་
1      ལི་      -      PROPN  -      -      -      -      -
2      ཀེ་      -      PROPN  -      -      -      -      -
3      ཁྱང་     -      PROPN  -      -      -      -      -
4      ལྷན་     -      ADP    -      -      -      -      -
5      ལྷན་ ལྷན་ ལྷན་ ལྷན་  VERB  -      -      -      -
6      -      -      PUNCT  -      -      -      -      -
```

Figure 6 Li Keqiang ལི་ཀེ་ཁྱང་, the name of the Chinese premier (2013–2023), is incorrectly tokenised into individual components.

## 5 Training a Basic spaCy Model for Modern Tibetan

This section describes the development of a modern Tibetan spaCy language model with the automatically created training data described in the previous section, comprising a pipeline that includes a POS-tagger, for integration into the iLCM. Given that the iLCM uses spaCy

<sup>26</sup> For the many ways in which foreign personal names and toponyms are rendered in Tibetan, see Erhard & Xiaoying 2025.

<sup>27</sup> In the current version of Modern Botok, we included more than 1,200 personal and place names found in newspapers from the 1950s and 1960s (Erhard *et al.* 2024). However, this did not include the majority of Chinese or foreign names found in the later newspapers that we included in the dataset.

version 3.2.6, we conducted training with version 3.2.1 to ensure compatibility.<sup>28</sup>

As an initial experiment, a spaCy language model, Tibetan for spaCy 1.1, was trained using texts in which Tibetan sentences were artificially segmented by white spaces with Botok, which had not yet been customised as described above, thereby avoiding directly integrating Botok into spaCy (Engels *et al.* 2023). It is important to note that during this training process, the model treated input sentences as if they were in the English language, with the language parameter in the configuration file set to English (`lang = 'en'`). Although this model enabled Tibetan language support on iLCM, performing tokenisation with the same accuracy as Botok, it has no POS-tagging capability, a prerequisite for many downstream NLP tasks.

### 5.1 *Training of the Modern Tibetan spaCy model*

In contrast, the approach discussed in this section involved training a spaCy model from scratch, i.e., setting the language parameter in the configuration file to multi-language (`lang = 'xx'`) and integrating Botok into the training process.

However, relying solely on Botok proved insufficient for developing a functional spaCy model for modern Tibetan. Although returning no error messages during training, the resulting spaCy model often misassigned POS-tags in an apparently random manner. This indicated the need to adjust the spaCy model's training configuration further.

Only through extensive experimentation we identified two critical configuration modifications that were necessary to handle the Tibetan data: (1) setting the pipeline to [`"tok2vec", "morphologizer"`] instead of [`"tok2vec", "tagger"`], and (2) adding the phrase `"SpaceAfter=No"`

---

<sup>28</sup> Note that the iLCM requires spaCy version 3.2.6, while the newest version was 3.8.2 at the time of writing. As the computational demands of model training exceeded the capabilities of a personal laptop, the models for Divergent Discourses were trained on the high-performance computing cluster maintained by the Scientific Computing team at Leipzig University (<https://www.sc.uni-leipzig.de/>).

in the MISC column of the CoNLL-U file.<sup>29</sup> We adopted this approach based on observed conventions in other language-specific CoNLL-U files, while additional settings could possibly further enhance the model’s performance.

```

# sent_id = 1
# text = ། ལྷ་ མེ་
1      །          །          PUNCT  _      _      0      root  _      SpaceAfter=No
2      ལྷ་ མེ་    ལྷ་ མེ་    NOUN   _      _      0      root  _      SpaceAfter=No

# sent_id = 2
# text = རྟེ་ ཡང་ ལྷ་ རྫོང་
1      རྟེ་ ཡང་    རྟེ་ ཡང་    ADV   _      _      0      root  _      SpaceAfter=No
2      ལྷ་ རྫོང་    ལྷ་ རྫོང་    NOUN  _      _      0      root  _      SpaceAfter=No

# sent_id = 3
# text = ལྷིན་ ལྷུ་ འགའ་ རྫོབ་ དང་ །
1      ལྷིན་        ལྷིན་        NOUN   _      _      0      root  _      SpaceAfter=No
2      ལྷུ་          ལྷུ་          ADP    _      _      0      root  _      SpaceAfter=No
3      འགའ་ རྫོབ་    འགའ་ རྫོབ་    NOUN  _      _      0      root  _      SpaceAfter=No
4      དང་          དང་          CCONJ  _      _      0      root  _      SpaceAfter=No
5      །            །            PUNCT  _      _      0      root  _      SpaceAfter=No

# sent_id = 4
# text = འབྲེ་ ལྷིན་ འདྲེ་ ལྷེགས།
1      འབྲེ་ ལྷིན་    འབྲེ་ ལྷིན་    NOUN   _      _      0      root  _      SpaceAfter=No
2      འདྲེ་ ལྷེགས།  འདྲེ་ ལྷེགས།  NOUN   _      _      0      root  _      SpaceAfter=No
3      །            །            PUNCT  _      _      0      root  _      SpaceAfter=No

# sent_id = 5
# text = མངས་ ལྷམ་ འཛོམ་ ལྷན་ འདམ་ ལྷིན་ འདེ་ འགྲན་ ལ་ ལྲང་ ལྷམ་ ལྲང་ ལྷ་ ལྷ་ འོང་ ལྷམ་ ལ་ ལྷིག་ ལྷུངས་ འོད།
1      མངས་ ལྷམ་    མངས་ ལྷམ་    NOUN   _      _      0      root  _      SpaceAfter=No
2      འཛོམ་ ལྷན་  འཛོམ་ ལྷན་  NOUN   _      NOUN   0      root  _      SpaceAfter=No
3      ལྷིན་        ལྷིན་        ADP    _      _      0      root  _      SpaceAfter=No
4      འ་          འ་          PRON   _      _      0      root  _      SpaceAfter=No
5      འདེ་        འདེ་        PRON   _      _      0      root  _      SpaceAfter=No
6      འགྲན་ ལ་    འགྲན་ ལ་    NOUN   _      _      0      root  _      SpaceAfter=No
7      ལྲང་        ལྲང་        NOUN   _      _      0      root  _      SpaceAfter=No
8      ལྷམ་        ལྷམ་        ADP    _      _      0      root  _      SpaceAfter=No
9      ལྲང་        ལྲང་        NOUN   _      _      0      root  _      SpaceAfter=No
10     ལྷ་          ལྷ་          ADP    _      _      0      root  _      SpaceAfter=No
11     ལྷ་          ལྷ་          VERB   _      _      0      root  _      SpaceAfter=No
12     འོང་        འོང་        VERB   _      _      0      root  _      SpaceAfter=No
13     ལྷམ་ ལ་    ལྷམ་ ལ་    PRON   _      _      0      root  _      SpaceAfter=No
14     ལྷིག་        ལྷིག་        DET    _      _      0      root  _      SpaceAfter=No
15     ལྷུངས་        ལྷུངས་        VERB   _      _      0      root  _      SpaceAfter=No
16     འོད་        འོད་        AUX    _      _      0      root  _      SpaceAfter=No
17     །            །            PUNCT  _      _      0      root  _      SpaceAfter=No

"train.conllu" 675979L, 39335983B

```

Figure 7 The final version of the CoNLL-U file produced by Gemini Pro 1.5.

Additionally, as the iLCM requires either a sentencizer or dependency parser to recognise sentence boundaries,<sup>30</sup> a post-pro-

<sup>29</sup> For a detailed explanation, refer to our discussion in the spaCy forum: <https://github.com/explosion/spaCy/discussions/13549> (accessed January 15, 2025).

<sup>30</sup> This is mainly to avoid errors such as: “ValueError: [E030] Sentence boundaries unset. You can add the 'sentencizer' component to the pipeline with:

cessing step was implemented. In this step, the HEAD and DEPREL fields in the CoNLL-U file were automatically populated with the placeholder values “root” and “0” respectively, allowing the training of a parser to function as a sentencizer. Despite such challenges during training with spaCy, we successfully developed a basic spaCy model for modern Tibetan (Kyogoku *et al.* 2024c).

## 5.2 Evaluation of the Modern Tibetan spaCy model

The accuracy scores (correct predictions / all predictions) of the Modern Tibetan spaCy, compared with the experimental Tibetan for spaCy 1.1 (based on an English language model) and Modern Botok, i.e., the customised Botok, are as presented in Table 1.<sup>31</sup>

Table 1 Evaluation table for tokenisation and POS-tagging.

<i>Method</i>	<b>Tokenisation</b>	<b>POS-tagging</b>
<i>Tibetan for spaCy 1.1</i>	0.931 (1500/1611)	N/A
<i>Modern Botok</i>	0.957 (1510/1578)	N/A
<i>Modern Tibetan spaCy</i>	0.924 (1394/1509)	0.872 (1316/1509)

All three methods demonstrate strong performance in terms of tokenisation. Tibetan for spaCy 1.1, which requires input texts with words separated by white spaces, tokenises texts precisely as the default, uncustomised Botok (see Appendix C). Notably, the tokenisation scores for both Modern Botok (see Appendix D) and

---

``nlp.add_pipe('sentencizer')`. Alternatively, add the dependency parser or sentence recogniser or set sentence boundaries by setting `doc[i].is_sent_start`."`

<sup>31</sup> In our evaluation standard, the tokenisation of compounds, such as ལྷོ་ལོ་ལོ་ ("outside and inside") and ལྷོ་ལོ་ལོ་ ("biggest"), into their constituent parts — ལྷོ་ ("outside") and ལོ་ ("inside"), or ལྷོ་ ("big") and ལོ་ ("most") — is permissible, alongside their undivided forms. On the other hand, if personal or place names are fragmented into meaningless components, such as ལྷོ་ལོ་ལོ་ (Xi Jinping 习近平) being split into ལྷོ་, ལོ་, and ལོ་, all syllables constituting the compound are considered erroneous in the evaluation. This evaluation uses the same dataset as described in Appendix B, comprising of 100 randomly selected sentences (see also Section 4.1).



Modern Tibetan spaCy (see Appendix E) differ. At the same time, they should be identical, in theory, given that the tokenisation in both cases is based on Botok. Still, the score for Modern Botok is slightly higher than that of the Modern Tibetan spaCy.

The relatively weaker performance of Modern Tibetan spaCy points to a limitation associated with using spaCy to process Tibetan language. First, since Tibetan texts have no word segmentation, spaCy requires the integration of the external tokeniser Botok, even though tokenised training data is provided in the CoNLL-U file. Since the CoNLL-U file generated by Gemini Pro 1.5 reflects the tokenisation results of Botok, presumably, the weakness is introduced in spaCy's training process.<sup>32</sup> Nevertheless, it is worth highlighting the high POS-tagging score achieved by Modern Tibetan spaCy.<sup>33</sup>

## 6 Conclusion

By addressing the unique syntactic features of Tibetan and overcoming the challenges posed by the scarcity of annotated corpora of modern Tibetan texts, a Modern Tibetan spaCy language model could be trained. This study demonstrated the potential of LLMs such as Gemini Pro to generate training data for low-resourced languages automatically. Divergent Discourses accordingly produced an automatically POS-tagged corpus of modern Tibetan within a relatively short time, facilitating the successful training of a spaCy language model. Existing bugs will likely be fixed by either improving Botok's tokenisation (e.g., by employing a fine-tuned Tibetan LLM) or

---

<sup>32</sup> Notably, this does not apply to our experimental Tibetan for spaCy 1.1. model, which was trained on Tibetan text segmented by white spaces.

<sup>33</sup> POS-tags associated with tokens evaluated as correct are assessed accordingly. Conversely, POS tags associated with tokens identified as erroneous are generally considered erroneous. As a result, the accuracy score for POS-tagging is inherently lower than that for tokenisation. When the number of correct predictions is divided by the total number of tokens evaluated as correct, the resulting accuracy score is 0.976 (1316/1394).

improving the POS tagging capabilities of LLMs such as Gemini, Claude, or ChatGPT.

The current Modern Tibetan spaCy model allows researchers to perform NLP analysis of Tibetan language materials with corpus analysis tools dependent on spaCy. The availability of a Tibetan language model for spaCy, an industry standard and openly accessible NLP platform, represents a significant step forward in enhancing digital accessibility and advancing linguistic research on Tibetan language textual sources. Although this research is an ongoing effort and future work will focus on enhancing and fine-tuning, particularly on enabling NER functionality for the Modern Tibetan spaCy model, the authors hope that the presented approach and workflow can function as a starting point for similar research on other under- and low-resourced languages.

### Bibliography

Barnett, Robert, Nathan W. Hill, Hildegard Diemberger, and Tsering Samdrup

"Named-Entity Recognition for Modern Tibetan Newspapers: Tagset, Guidelines and Training Data," *Zenodo*, 2021. [doi:10.5281/zenodo.4536516](https://doi.org/10.5281/zenodo.4536516).

Engels, James, Franz Xaver Erhard, Robert Barnett, and Nathan W. Hill  
"Tibetan for Spacy 1.1," *Zenodo*, 2023. [doi:10.5281/zenodo.10120779](https://doi.org/10.5281/zenodo.10120779)

Erhard, Franz Xaver

"Text and Layout Recognition for Tibetan Newspapers with Transkribus," *Revue d'Etudes Tibétaines* (74), 2025, pp. 128–171.

Erhard, Franz Xaver, Yuki Kyogoku, Robert Barnett, and Nathan W. Hill

“Modern-Botok: Custom Dictionary for Modern Tibetan (v0.1),” *Zenodo*, 2024. [doi:10.5281/zenodo.14034747](https://doi.org/10.5281/zenodo.14034747).

Erhard, Franz Xaver, and Xiaoying 笑影

“Foreign Names and Places in Tibetan Newspapers of the 1950s and 1960s,” *Revue d’Etudes Tibétaines* (74), 2025, pp. 172–186.

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

“Toponyms and Anthroponyms from Tibetan-language Newspapers of the 1950s and 1960s: Three Name Lists (v1.0),” [Data set]. *Zenodo*, 2024. [doi:10.5281/zenodo.14289491](https://doi.org/10.5281/zenodo.14289491)

Dakpa, Jamyang, Tashi Dhondup, Yeshe Jigme Gangne, Edward Garrett, Marieke Meelen, and Sonam Wangyal.

“Modern Tibetan corpus annotated for verb-argument dependency relations (v1.0),” [Data set]. *Zenodo*, (2021). [doi:10.5281/zenodo.4727129](https://doi.org/10.5281/zenodo.4727129).

Faggionato, Christian, and Edward Garrett.

“Constraint grammars for Tibetan language processing,” *NoDaLiDa 2019 Workshop on Constraint Grammar-Methods, Tools and Applications, Linköping Electronic Conference Proceedings* 168 (3), 2019. Available online at <http://www.ep.liu.se/ecp/168/003/ecp19168003.pdf> (accessed January 24, 2025).

Faggionato, Christian, Edward Garrett, Nathan W. Hill, Samyo Rode, Nikolai Solmsdorf, and Sonam Wangyal.

“Classical Tibetan Corpus Annotated for Verb-Argument Dependency Relations (v1.0),” *Zenodo*, 2021. [doi:10.5281/zenodo.4727108](https://doi.org/10.5281/zenodo.4727108).

Hackett, Paul G.

“Automatic Segmentation and Part-Of-Speech Tagging For Tibetan: A First Step Towards Machine Translation.” In *Proceedings of the 9th Seminar of the International Association for Tibetan Studies*, 2000, pp. 1-18. Available online at <http://hdl.handle.net/10022/AC:P:10471> (accessed on December 18, 2024).

Garrett, Edward, and Nathan W. Hill.

“Constituent order in the Tibetan noun phrase,” *SOAS Working Papers in Linguistics* 17, 2015, pp. 35–48.

Huang, Haoyang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei

“Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting,” *arXiv preprint*, 2023. [doi:10.48550/arXiv.2305.07004](https://doi.org/10.48550/arXiv.2305.07004).

Kyogoku, Yuki, Franz Xaver Erhard, Robert Barnett, and Nathan W. Hill

“TibNorm - Normaliser for Tibetan (Version v1),” *Zenodo*, 2024a. [doi:10.5281/zenodo.10815272](https://doi.org/10.5281/zenodo.10815272)

“Diverge-Gemini POS-tagged Corpus of Modern Tibetan (1.0),” [Data set]. *Zenodo*, 2024b. [doi:10.5281/zenodo.14447192](https://doi.org/10.5281/zenodo.14447192).

“Basic Modern Tibetan SpaCy Model,” *Zenodo*, 2024c. [doi:10.5281/zenodo.10806456](https://doi.org/10.5281/zenodo.10806456).

Li, Yan, Xiaomin Li, Yiru Wang, Hui Lü, Fenfang Li, and La Duo

“Character-based Joint Word Segmentation and Part-of-Speech Tagging for Tibetan Based on Deep Learning,” *Transactions on Asian and Low-Resource Language Information Processing*, 21 (5), Article 95, 2022, pp. 1-15, [doi:10.1145/3511600](https://doi.org/10.1145/3511600).

Lv, Hui, Chi Pu, La Duo, Yan Li, Qingguo Zhou, and Jun Shen

“T-LLaMA: a Tibetan large language model based on

LLaMA2." *Complex and Intelligent Systems* 11 (1), 2025. [doi:10.1007/s40747-024-01641-7](https://doi.org/10.1007/s40747-024-01641-7).

Meelen, Marieke, and Nathan W. Hill

"Segmenting and POS tagging Classical Tibetan using a memory-based tagger," *Himalayan Linguistics*, 16 (2), 2017.

Meelen, Marieke, Élie Roux, and Nathan W. Hill

"Optimisation of the Largest Annotated Tibetan Corpus Combining Rule-Based, Memory-Based, and Deep-Learning Methods," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 20 (1), 2021, pp. 1-11. [doi:10.1145/3409488](https://doi.org/10.1145/3409488).

Niekler, Andreas, Gregor Wiedemann, and Gerhard Heyer

"Leipzig Corpus Miner: A Text Mining Infrastructure for Qualitative Data Analysis." In *Terminology and Knowledge Engineering*, 2014. Available online: <https://hal.archives-ouvertes.fr/hal-01005878> (accessed on December 18, 2024).

Niekler, Andreas; Christian Kahmann, Manuel Burghardt, and Gerhard Heyer

"The interactive Leipzig Corpus Miner. An extensible and adaptable text analysis tool for content analysis." In: *Publizistik* 68 (2-3), 2023, pp. 325–354. [doi:10.1007/s11616-023-00809-4](https://doi.org/10.1007/s11616-023-00809-4).

Sivarajkumar, Sonish, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang.

"An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study," *JMIR Medical Informatics* 12, 2024.

van der Sloot, Ko, and Maarten van Gompel

"LanguageMachines/mbt (v3.11)," *Zenodo*, 2024. [doi:10.5281/zenodo.14500870](https://doi.org/10.5281/zenodo.14500870)

Xiangxiu, Cairang, Nuo Qun, Nuobu Renqing, Trashi Nyima and Qijun Zhao.

“Research on Tibetan Part-of-Speech Tagging Based on Transformer,” *3rd International Conference on Pattern Recognition and Machine Learning (PRML)*, Chengdu, 2022, pp. 315-320. [doi: 10.1109/PRML56267.2022.9882234](https://doi.org/10.1109/PRML56267.2022.9882234).

## Appendices

### Appendix A: Comparing Expert vs Botok POS-Tagging

	Expert		Botok		Botok passes	Botok errors
0	བཟློན་དོན་	NOUN	བཟློན་དོན་	OTHER	x	
1	གཙོ་བོ་	ADJ	གཙོ་བོ་	ADJ		
2	འི་	ADP	འི་	PART		
3	སློབ་གསོ་	NOUN	སློབ་གསོ་	OTHER	x	
4	ལག་	DET	ལག་	NOUN		
5	གཉེས་པ་	ADJ	གཉེས་པ་	NO_POS	x	
6	ཕྱི་ལོ་འཛུགས་	NOUN	ཕྱི་ལོ་འཛུགས་	NO_POS	x	
7	ཞིབ་ཚགས་	NOUN	ཞིབ་ཚགས་	NO_POS	x	
8	བྱས་	VERB	བྱས་	VERB		
9	ནས་	SCONJ	ནས་	PART		
10	ལག་བསྟར་	NOUN	ལག་བསྟར་	OTHER	x	
11	བྱ་རྒྱུ་	VERB	བྱ་	VERB		
12			རྒྱུ་	NOUN		x
13	ནི་	PART	ནི་	NO_POS	x	
14	མིག་སྣ་	NOUN	མིག་	NOUN		
15			སྣ་	ADV		x
16	འི་	ADP	འི་	PART		
17	བོད་ལྗོངས་	PROPN	བོད་ལྗོངས་	PROPN		
18	ཡོངས་	DET	ཡོངས་	DET		
19	ཀྱི་	ADP	ཀྱི་	PART		
20	ས་གནས་	NOUN	ས་གནས་	OTHER	x	
21	ལག་	NOUN	ལག་	NOUN		
22	དང་	CCONJ	དང་	NO_POS	x	
23	ཚན་པ་	NOUN	ཚན་པ་	OTHER	x	
24	ལག་	NOUN	ལག་	NOUN		
25	གི་	ADP	གི་	PART		

	Expert		Botok		Botok passes	Botok errors
26	གལ་ཚེ	NOUN	གལ་ཚེ	OTHER	x	
27	འི་	ADP	འི་	PART		
28	ལས་དོན་	NOUN	ལས་དོན་	OTHER	x	
29	ཞིག་	DET	ཞིག་	PART		
30	ཡིན	AUX	ཡིན	NO_POS	x	
31		PUNCT		PUNCT		
					13	2

*Appendix B Evaluation of Gemini (tokenisation and) POS-tagging*

# text = ཚོགས་འདུ་འི་ ཐོག་ སྤྱི་ཚབ་ མུའུ་ ཅི་ ཞེ་ ཅིན་ མིང་ གི་ དམག་དཔུང་ ལྷོ་བས་ ཚེ་ ར་ གཏོང་བ་ འི་ དགོངས་པ་ དང་ |

#	Token	Flag	POS	Flag
1	ཚོགས་འདུ	TRUE	NOUN	TRUE
2	འི་	TRUE	ADP	TRUE
3	ཐོག་	TRUE	ADP	TRUE
4	སྤྱི་ཚབ་མུའུ་ཅི་	FALSE	NOUN	TRUE
5	ཞེ་ཅིན་མིང་	TRUE	PROPN	TRUE
6	གི་	TRUE	ADP	TRUE
7	དམག་དཔུང་	TRUE	NOUN	TRUE
8	ལྷོ་བས་ཚེ་	FALSE	VERB	FALSE
9	གཏོང་བ་	TRUE	AUX	TRUE
10	འི་	TRUE	ADP	TRUE
11	དགོངས་པ་	TRUE	NOUN	TRUE
12	དང་	TRUE	CCONJ	TRUE
13		TRUE	PUNCT	TRUE



*Appendix C Evaluation of Tibetan for spaCy 1.1 (tibetan\_tib\_en\_ver1-0.0.1) tokenisation*

# text = ཚོགས་འདུ་ འི་ ཐོག་ སྤྱི་ཚུབ་ ཟུའུ་ ཅི་ ཞི་ ཅིན་ མིང་ ལྷི་ དམག་དཔུང་ ལྷོབས་ ཚེ ར་ གཏོང་བ འི་ དགོངས་པ་ དང་ །

#	Token	Flag
1	ཚོགས་འདུ	TRUE
2	འི་	TRUE
3	ཐོག་	TRUE
4	སྤྱི་ཚུབ་	TRUE
5	ཟུའུ་	FALSE
6	ཅི་	FALSE
7	ཞི་	FALSE
8	ཅིན་	FALSE
9	མིང་	FALSE
10	ལྷི་	TRUE
11	དམག་དཔུང་	TRUE
12	ལྷོབས་	TRUE
13	ཚེ	TRUE
14	ར་	TRUE
15	གཏོང་བ	TRUE
16	འི་	TRUE
17	དགོངས་པ་	TRUE
18	དང་	TRUE
19	།	TRUE

*Appendix D Evaluation of Modern Botok tokenisation*

# text = ཚོགས་འདུ་ འི་ ཐོག་ སྤྱི་ཚུབ་ ཟུའུ་ ཅི་ ཞི་ ཅིན་ མིང་ ལྷི་ དམག་དཔུང་ ལྷོབས་ ཚེ ར་ གཏོང་བ འི་ དགོངས་པ་ དང་ །

#	Token	Flag
1	ཚོགས་འདུ	TRUE
2	འི་	TRUE

#	Token	Flag
3	ཐོག་	TRUE
4	སྤྱི་ལུག་	TRUE
5	རྒྱ་ཅི་	TRUE
6	ཞི་ཅིན་མིང་	TRUE
7	གི་	TRUE
8	དམག་དཔུང་	TRUE
9	སྟོབས་ཆེར་	FALSE
10	གཏོང་བ	TRUE
11	འི་	TRUE
12	དགོངས་པ་	TRUE
13	དང་	TRUE
14		TRUE

*Appendix E Evaluation of Modern Tibetan spaCy (xx\_bo\_tagger-0.1.2) tokenisation and POS-tagging*

# text = ཚོགས་འདུ་འི་ཐོག་སྤྱི་ལུག་རྒྱ་ཅི་ཞི་ཅིན་མིང་གི་དམག་དཔུང་སྟོབས་ཆེར་གཏོང་བའི་དགོངས་པ་དང་།

#	Token	Flag	POS	Flag
1	ཚོགས་འདུ	TRUE	NOUN	TRUE
2	འི་	TRUE	ADP	TRUE
3	ཐོག་	TRUE	NOUN	TRUE
4	སྤྱི་ལུག་རྒྱ་ཅི་	FALSE	NOUN	TRUE
5	ཞི་ཅིན་མིང་	TRUE	NOUN	FALSE
6	གི་	TRUE	ADP	TRUE
7	དམག་དཔུང་	TRUE	NOUN	TRUE
8	སྟོབས་ཆེར་	FALSE	NOUN	FALSE
9	གཏོང་བ	TRUE	VERB	TRUE
10	འི་	TRUE	ADP	TRUE
11	དགོངས་པ་	TRUE	NOUN	TRUE
12	དང་	TRUE	CCONJ	TRUE
13		TRUE	PUNCT	TRUE

Appendix F Comparing POS tagging by three LLMs (September, 2024)

Text = བརྗོད་ཚོན་གཞི་མེད་སློབ་གསོ་ཁག་གཉིས་པ་སྤྲིན་འཇུག་ཞིབ་ཚུགས་བྱས་ནས་ལག་བསྟར་བྱ་རྒྱུ་ནི་མིག་སྡེ་བོད་ལྗོངས་ཡོངས་ཀྱི་ས་གནས་ལག་དང་ཚོན་པ་ལག་གི་གསལ་ཚེད་ལས་རྗོད་ཞིག་ཡིན།


Expert		ChatGPT4				ChatGPT4o				Claude 3.5				Gemini			
		Token same as expert Y/N	POS-tags	Token same as expert, POS tag is same (Y) or equivalent (Y~) to expert	Token same as expert Y/N	POS-tags	Token same as expert, POS tag is same (Y) or equivalent (Y~) to expert	Token same as expert Y/N	POS-tags	same as expert, POS tag is same (Y) or equivalent (Y~) to expert	Token same as expert Y/N	POS-tags	same as expert, POS tag is same (Y) or equivalent (Y~) to expert	Token same as expert Y/N	POS-tags		
བཟོད་ཚོན་	NOUN	N	NOUN		N	VERB		Y	NOUN	Y	Y	NOUN	Y				
		N	NOUN		N	NOUN											
གཞི་མེད་	ADJ	N	ADJECTIVE		N	ADJECTIVE		N	ADJECTIVE		N	ADJECTIVE					
		N	ADP		N	GEN. PART											
སློབ་གསོ་	NOUN	N	NOUN		N	NOUN		Y	NOUN	Y	Y	NOUN	Y				
		N	NOUN		N	NOUN											
ལག་	DET	Y	NOUN	N	Y	NOUN	N	Y	PLURAL	Y~	Y	NOUN	N				
		Y	ORD. NUMBER	Y~	Y	ADJECTIVE	Y	Y	NUMERAL	Y~	Y	NUMERAL	Y~				
སྤྲིན་འཇུག་	NOUN	N	VERB		N	VERB		Y	NOUN	Y	Y	VERB	N				
		N	VERB		N	VERB											
ཞིབ་ཚུགས་	NOUN	N	ADJECTIVE		N	ADJECTIVE		Y	ADVERB	Y~	Y	VERB	N				
		N	NOUN		N	NOUN											
བྱས་	VERB	Y	VERB	Y	Y	VERB	Y	N	VERB		N	CONJUNCTION					
		Y	ABL. PART	Y~	Y	ABL. PART	Y~										
ལག་བསྟར་	NOUN	N	NOUN		N	NOUN		Y	NOUN	Y	N	NOUN					
		N	VERB		N	VERB											
བྱ་	VERB	N	NOUN		N	VERB		Y	VERB NOMINAL	Y~							
		N	NOUN		N	NOUN											
ནི་	PART	Y	VERB. AUX PART	N	Y	NOM. PART	Y~	Y	TOPIC MARK.	Y~	Y	COPULA	N				
		N	NOUN		N	NOUN		N	ADJECTIVE		N	ADJECTIVE					
མིག་སྡེ་	NOUN	N	NOUN		N	NOUN											
		N	ADJECTIVE		N	NOUN											
མེད་	ADP	Y	GEN. PART	Y~	Y	GEN. PART	Y~										
		N	PROPN		N	NOUN		Y	NOUN	Y~	Y	PROPER NOUN	Y				
མེད་ལས་	NOUN	N	NOUN		N	NOUN											
		N	NOUN		N	NOUN											
ཡོངས་	DET	Y	NOUN	N	Y	ADJECTIVE	N	Y	ADVERB	N	N	ADJECTIVE					
		Y	GEN. PART	Y	Y	GEN. PART	Y	Y	GEN. PART	Y							
ཀྱི་	ADP	N	NOUN		N	NOUN		Y	NOUN	Y	Y	NOUN	Y				
		N	NOUN		N	NOUN											
ལག་	DET	Y	NOUN	N	Y	NOUN	N	Y	PLUR. MARK.	Y~	Y	NOUN	N				
		Y	CCONJ	Y	Y	CONJ.	Y	Y	CONJ.	Y~	Y	CONJ.	Y				
དང་	CCONJ	Y	CONJ.	Y	Y	CONJ.	Y	Y	CONJ.	Y~	Y	CONJ.	Y				
		Y	NOUN	Y	Y	NOUN	Y	Y	NOUN	Y~	Y	NOUN	Y				
ཚོན་པ་	NOUN	Y	NOUN	Y	Y	NOUN	Y	Y	NOUN	Y~	Y	NOUN	Y				
		Y	DET	Y	NOUN	N	Y	PLUR. MARK.	Y~	Y	NOUN	N					
ལག་	DET	Y	NOUN	N	Y	NOUN	N	Y	PLUR. MARK.	Y~	Y	NOUN	N				
		Y	ADP	Y	GEN. PART	Y	Y	GEN. PART	Y~	Y	GEN. PART	Y~	Y	GEN. PART	Y~		
གསལ་ཚེད་	NOUN	N	NOUN		N	NOUN		N	ADJECTIVE		N	ADJECTIVE					
		N	ADP		N	ADJECTIVE											
ཞི་	ADP	N	ADJECTIVE		N	ADJECTIVE											
		Y	GEN. PART	Y~													
ལས་རྗོད་	NOUN	N	NOUN		N	NOUN		Y	NOUN	Y	Y	NOUN	Y				
		N	NOUN		N	NOUN											
ཞིག་	DET	Y	DEFIN. ART.	Y~	Y	NOMINAL	N	Y	INDEF. ART.	Y~	Y	DETERM.	Y				
		Y	AUX	Y	Y	VERB	Y~	Y	VERB	Y~	Y	COPULA	Y~				
		Tags	Tags	Tags	Tags	Tags	Tags	Tags	Tags	Tags	Tags	Tags	Tags	Tags	Tags		
		Y=15	Y~=10	Y=16		Y~=11	Y=21		Y=20	Y=17		Y=11					
		N=24	N=5	N=24		N=5	N=4		N=1	N=6		N=6					
Error score relative to expert		ChatGPT4				ChatGPT4o				Claude 3.5				Gemini			
	Tags	N=24		N=5	N=24		N=5		N=1		N=6		N=6				
Errors as %		61,5%		33,3%	60,0%		31,3%		16,0%		4,8%		26,1%		35,3%		

It must be noted that the above comparison, although providing a good overview of the tokenisation and POS tagging capabilities of the tested LLMs, is a snapshot of September 2024. The situation now, at the time of writing a few months later, has changed significantly. Most models have shown improved capabilities in translating and dealing with Tibetan. The situation was also different in June 2024 – when most of the training data for the Divergent Discourses project was created. Then Gemini Pro 1.5 performed best on Tibetan POS tagging tasks. While we had our script creating annotated training data with Gemini Pro 1.5 using Google's cloud API, Anthropic released its newest Claude 3.5 Sonnet model, a powerful alternative for Tibetan language tasks, on June 20, 2024, and an upgrade on October 22, 2024. Although we could not use Claude 3.5 to create training data, we wanted to include it in this comparison to show its good performance in Tibetan language-related tasks.



# Religious Policy in the TAR, 2014–24: Topic Modelling a Tibetan-Language Corpus with BERTopic

Ronald Schwartz (Memorial University of Newfoundland)  
and  
Robert Barnett (SOAS University of London)

he study of discourses, in the sense of narratives or themes, is essentially an historical project: a discourse generally has a life-span of some sort, emerging at a certain time, reaching a degree of prominence or pervasiveness, and then, in most cases, fading away or shifting into some other form. In this paper, we look at a set of computational tools that can be developed for tracing the epihistorical footprints left by certain discourses and discuss their use for the analysis and detection of such histories in modern Tibetan texts. Ultimately, these tools will be useable with any Tibetan texts that have been digitised, such as the newspapers from the 1950s and 1960s that are the target of the Divergent Discourses project.<sup>1</sup> For the initial development and testing of these tools, however, we used a set of texts in modern Tibetan that are “born-digital” – that is, they are already available in a digital format and, therefore, do not need to be photographed, scanned or machine-read.

The texts that we used for this study are articles taken from the premier Tibetan-language newspaper in the People’s Republic of

---

<sup>1</sup> The project received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (<https://gepris.dfg.de/gepris/projekt/508232945?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/>.

China (PRC), *Tibet Daily*, known in Chinese as *Xizangribao* (西藏日报) and in Tibetan as the *Bod ljongs nyin re'i tshags par* (བོད་ལྗོངས་ཉིན་རེའི་ཚགས་པར་). *Tibet Daily* is not a newspaper in the normal sense of the word, where the principal purpose is to disseminate news. Rather, it is what might elsewhere be known as a gazetteer, in that it serves primarily to publicise decisions, projects, goals and opinions held by its proprietary institution: the Committee of the Tibet Autonomous Region (TAR) Branch of the Chinese Communist Party (CCP). As a result, we can, as analysts, assume that in most cases, a “discourse” or topic found in the pages of *Tibet Daily* will reflect not one or other commonly held opinion or narrative circulating organically within society, but a project, policy, opinion or goal that the TAR branch of the CCP intends at that time to publicise or implement.

In most cases, the presence of such a discourse in the columns of *Tibet Daily* will indicate that the CCP has embarked upon an organised effort to disseminate a specific belief or practice among the public. These efforts will reflect the initiation of a policy or political program, which at times will take the particularly intense form of political or social mobilisation known in Chinese as a campaign (运动 *yundong*; Tib. *las 'gul*), when Party officials, government workers and Party activists will be sent throughout the region to achieve the goals of that campaign. Consequently, a computational tool that can trace the historical arc of a discourse or topic can, when applied to a publication such as the *Tibet Daily*, be used to identify the beginnings, peaks and endings of the political campaigns, drives, policies and official opinions that shape, or attempt to shape, much of public and private life in contemporary Tibet. This paper describes our development of a dynamic topic modelling tool for modern Tibetan texts that can help analysts trace the rise and fall of such campaigns, policies and opinions in Tibet over time.

### 1 Creating the *Tibet Daily* corpus

*Tibet Daily* began publication in 1956. Nearly a year earlier, in October 1955, China's then leader, Mao Zedong (1893–1976), had sent a

message to Zhang Jingwu (张经武, 1906–1971), the Chinese government’s leading representative in Tibet, endorsing the decision to produce a newspaper in Tibet that would serve as the official organ of the CCP in Tibet. Reportedly, Mao Zedong wrote to Zhang:

"When running a newspaper in a minority area, the first thing to do is to run a newspaper in the minority language ... Unlike Qinghai, Tibet should not have a newspaper which is [bilingual] in Tibetan and Chinese, but [it should have one that is only] in Tibetan. The name of the newspaper and how it should be run should be discussed with the Tibetan authorities, who should decide, and we should not take over the running of the newspaper."<sup>2</sup>

This advice, emblematic of the concessional, co-optive approach of the CCP leadership towards central Tibetans at the time, was taken up by Zhang and his colleagues, at least on the surface: the new paper was published in two separate editions, one in Chinese and another in Tibetan, and Tibetan leaders were included in the decision as to its name. The paper rapidly became a major institution in the region, representing a core function of the new Chinese administration in Tibet, with a staff of 350 people by the early 2000s (Zhang 2004: 141). The political importance of the paper was underlined by the fact that it had to be defended by its own 60-strong militia unit during the uprising of March 1959, when it came under sustained, but ultimately unsuccessful, attack from Tibetan rebels (Zhao 1987: 179). Much the same happened in July 1966, when leftist activists besieged the *Tibet Daily* compound, this time in the name of Chairman Mao rather than

---

<sup>2</sup> “在少数民族地区办报，首先应办少数民族文字的报。” “西藏与青海不同，不要藏汉两文合版，要办藏文报。报纸用什么名字和怎样办好，应同西藏地方商量，由他们决定，我们不要包办。” (Dangdai Zhongguo congshu bianjibu [Editorial Board of the Series on Contemporary China], 当代中国的西藏 [Tibet in Contemporary China], vol. 2, Beijing: Contemporary China Press, 1991, p. 435). See <https://www.google.co.uk/books/edition/当代中国的西藏/UYm6AAAAIAAI> (accessed January 15, 2025); see also “*Tibet Daily*”, unsigned entry primarily authored by user “TinaLees-Jones”, Wikipedia, [https://en.wikipedia.org/wiki/Tibet\\_Daily](https://en.wikipedia.org/wiki/Tibet_Daily) (accessed January 15, 2025).

Tibetan independence (Goldstein *et al.* 2009: 27–30; Tsering Woeser 2020: 372).

Although Mao is said to have emphasised the importance of the Tibetan-language edition of *Tibet Daily*, the Chinese-language version has come to have, and probably always had, a dominant role. By the mid-1990s, the print run for the Chinese edition was nearly 30,000, while the Tibetan edition reached 20,000; the number of contributors to the Chinese edition by that time was over 2,300, about twice the number who had written for the paper in Tibetan (Hartley 2005: 248). In addition, our research has found that, at least since 2014, there are substantially more articles in the Chinese-language edition of *Tibet Daily* than in the Tibetan-language edition. On average there are roughly twice as many articles in a given year in the Chinese edition compared to the Tibetan. Thus, in 2020, for instance, there were 6,295 articles in the Tibetan edition and 12,146 in the Chinese edition. More significantly, most Tibetan articles appeared to be translations of Chinese articles. We have not explored the differences between the two editions, but it is clear that the content we are primarily interested in — announcements of new policies and their implementation, important political meetings, speeches by leaders, slogans and political campaigns — seem to appear in the Chinese edition and then, usually a day or more later, are mirrored in the Tibetan version.

However, the Tibetan-language edition of *Tibet Daily* has played a significant role in promoting the vernacularisation of the language, according to Hartley (2005). Following the death of Mao and the end of the Cultural Revolution, it also made some contributions to the development of modern Tibetan literature, particularly by adding a literary column called *Smyug bsar* [New Pen] and, from the early 1980s onwards, publishing occasional short stories by Tibetan writers (Hartley 2005: 248).

For 40 years, the two editions of *Tibet Daily* were produced on paper, mainly for distribution in government offices and similar institutions. From the early 21<sup>st</sup> century, however, *Tibet Daily* began to appear online as well as in print. Currently, it is available in several online versions, with its main Chinese-language hub at <https://www.xzxw.com/> (accessed January 15, 2025; the site name is an acronym of



*Xizangxinwen*, “Tibet news”). This web portal aggregates multiple articles on different subjects published on different dates from various newspapers produced by the authorities in the TAR. These articles are accessed through links on pages that facilitate recursive navigation. This site has subsidiary pages which carry Tibetan-language and English-language versions of articles from *Tibet Daily* and other official publications, again in an aggregated format.

However, since 2008 *Tibet Daily* has also been available online in an e-paper format. This format reproduces the look of the printed edition, with a photographic image (actually, a PDF) of each page on the left of the screen and a text version of each article on the right. This static website has a flat structure, with content organised as a single daily edition with no recursive access to previous editions. The individual articles are HTML files organised chronologically by date, and thus it is possible to assemble a corpus covering the entire timespan of the e-paper website to the present. The Tibetan-language edition is accessed separately at a site with the root name <https://e.xzxw.com/xzrbzw/> (for Xizangribao Zangwen, “Tibetan-language Tibet Daily”), followed by the date of that day’s issue.

Articles from the Chinese-language edition of *Tibet Daily* are available in HTML format at <https://e.xzxw.com/xzrb/> (followed by the date) from 2008 onwards. However, the Tibetan-language site only has articles in HTML format from 2014. Before that, each issue of the e-paper is displayed as single whole pages in PDF format and thus is not easily recoverable as text. For the purpose of this study, we will therefore use the Tibetan-language edition of *Tibet Daily* from 2014 to 2024.

The articles we collected are all in HTML format and use a Unicode-compliant font for Tibetan characters. To facilitate analysis using the research tools developed for this study, the content of the articles was extracted using tags embedded in the HTML files. The paragraphs within the articles have been individually tagged, enabling the creation of a CSV file for each year of the corpus, with a row for every paragraph, along with metadata (article titles, dates, filename of original article). These CSV files comprise the corpus for research.

## 2 Semantic Searching with Vector Embeddings

This paper will demonstrate the use of tools that employ vector embeddings derived from transformer-based large language models (LLMs) to analyse Tibetan texts. Embeddings are numerical encodings that locate lengths of texts (phrase, sentences, paragraphs) in a high-dimensional vector space where semantically similar texts are close together in the vector space. Embeddings capture contextual meaning rather than mere word co-occurrences, providing a richer representation of language and meanings. There are a number of embedding models available for high-resource languages such as English or Chinese, but Tibetan is a relatively low-resource language where training has necessarily been limited to relatively small datasets. Ideally, multilingual embedding models will locate texts with the same or similar meanings in two or more different languages close to each other within the vector space.<sup>3</sup>

After examining several multilingual models, we found one that performs well with modern newspaper Tibetan — the version 2.0

---

<sup>3</sup> There has been relatively little work to date using vector embeddings with Tibetan-language texts. Meelen (2022) reports using FastText (<https://fasttext.cc/>, accessed January 15, 2025) to generate word embeddings for classical Tibetan. Sabbagh (2023) uses the LASER multilingual sentence encoder from Facebook (<https://github.com/facebookresearch/>, accessed January 15, 2025) to align Tibetan translations of English language sentences. Neither of these models are transformer-based. Two transformer-based embedding models have been developed by researchers in the PRC using the BERT model — Tibetan-BERT from a team at Tibet University ([https://huggingface.co/UTibetNLP/tibetan\\_bert](https://huggingface.co/UTibetNLP/tibetan_bert), accessed January 15, 2025) and TiBERT from a team at Minzu University (<https://huggingface.co/CMLI-NLP/TiBERT>, accessed January 15, 2025). Both of these models are designed and trained for downstream tasks of text classification. BGE-m3 from the Beijing Academy of Artificial Intelligence is an open-source multilingual embedding model, designed for information retrieval applications, that offers Tibetan embeddings (<https://huggingface.co/BAAI/bge-m3>, accessed January 15, 2025). Amazon Web Services Titan text embeddings v2 model also includes Tibetan embeddings (<https://docs.aws.amazon.com/bedrock/latest/userguide/titan-embedding-models.html>, accessed January 15, 2025). Neither of these performed adequately for purposes of semantic searching and topic modelling. The newer Cohere version 3.0 model, which specialises in information retrieval tasks, also did not perform as well as version 2.0 (see Engels *et al.* 2025).

multilingual model from Cohere (<https://cohere.com/>, accessed January 15, 2025; see Engels & Barnett 2025 in this volume). The embeddings are optimised for multilingual text understanding and accessible through an API. We found that they also perform well with cross-lingual queries in Tibetan, Chinese, and English. This model has a 256-token context limit, which requires that the paragraphs must be “chunked” into phrases that fall within this limit before generating the embeddings. A routine is implemented to “chunk” or split the paragraphs at the last (!) *shad* (the Tibetan punctuation mark most often used to mark the end of a sentence or phrase) within 256 tokens. Vector embeddings are generated and aligned for every chunked paragraph using the Cohere API.<sup>4</sup>

Semantic searching is the basic tool for investigating the corpus. Topic modelling, which we describe later, relies on the same vector embeddings as semantic searching and applies the same measure of similarity. Using the Cohere embeddings we can search throughout the entire corpus of *Tibet Daily* from 2014 to 2024 for chunks of text (and corresponding paragraphs) that are semantically similar to our query text. The query can be just a few words or a phrase, but for exploring the corpus and identifying topics and themes, it is more effective to include a chunk of text drawn from the corpus that is representative of the content being searched for rather than an individual word or short string. The semantic search program (written in Python) calls the Cohere API to generate an embedding for the query and then compares the numerical encoding of the query with the previously generated encodings of every chunk in the entire corpus. The results are displayed in descending order of similarity. If

---

<sup>4</sup> The Cohere embedding model uses a WordPiece tokeniser, which segments the texts into single Tibetan syllables and uses ## for subword units. The Tibetan *tsheg* is also treated as a token. An example of a tokenised text: [ '[CLS]', 'མ', '##ར', ' ', 'ཅ', '##ུན', ' ', '།', '##ེ', '##ེང', ' ', 'གིས', ' ', 'ནན', ' ', 'བཤད', ' ', 'གནང', ' ', 'དོན', '།', 'མི', ' ', 'མ', ' ', 'ལག', ' ', 'དང', ' ', 'ཟེ', ' ', 'ཚོན', ' ', 'ལག', ' ', 'གིས', ' ', 'དག', ' ', 'ཟེ', '##ར', ' ', 'ལ', ' ', 'དམ', ' ', 'འཛིན', ' ', 'ཡག', ' ', 'མོ', ' ', 'ཅུ', ' ', 'ཞུ', ' ', 'ནི', ' ', 'རང', ' ', 'ལ', '##ལ', '##ེའི', '##འི', ' ', 'ལས', ' ', 'འགན', ' ', 'ཡིན', ' ', 'མ', ' ', 'དང', ' ', '།', 'དག', ' ', 'ཟེ', '##ར', ' ', 'ལ', ' ', 'དམ', ' ', 'འཛིན', ' ', 'མ', ' ', 'བྱས', ' ', 'མ', ' ', 'ནི', ' ', 'འགན', ' ', 'ཤོང', ' ', 'ཡིན', ' ', 'མ', ' ', 'དག', ' ', 'ཟེ', '##ར', ' ', 'ཡག', ' ', 'མོ', ' ', 'མ', ' ', 'བྱས', ' ', 'མ', ' ', 'ནི', ' ', 'འགན', ' ', 'ལ', '##ཇ', '##ོལ', ' ', 'ཡིན', ' ', 'མ', ' ', 'བཅས', ' ', 'ཉི', ' ', 'ལུ', ' ', 'གིས', ' ', 'བཤད', ' ', 'མོ', ' ', 'བརྟགས', ' ', 'ནས', '།', '[SEP]' ].

the query is itself a chunk from the corpus, it will be displayed first in the list of results and will have the highest similarity value. The program allows the user to specify the number of hits to return. For our research purposes, we might ask it initially to return as many as fifty or one hundred hits in descending order of similarity to get a sense of the scope of a query, as we are interested not just in the few top-most similar hits, but in exploring similar examples of discourse in many paragraphs in many articles published at different times.

Table 1 is an example of the results from a query using semantic search. Just the first four rows from a search are displayed here to illustrate the use of the search tool. In the first row, the query (a selection of text from the corpus) returns itself. The next three rows are paragraphs in descending order of similarity to the query. In this example, the query is used to locate material in the corpus that mentions implementing the “four standards” drive in monasteries and nunneries (one of the topics that will be discovered through topic modelling). Having found an instance of the “four standards” in one context in the corpus, semantic search makes it possible to find other contexts as well: #1 refers to a high-level meeting of officials and religious leaders in which the drive is discussed; #2 refers to the implementation of “four standards” education in Shigatse; and in #3, Ding Yexian, a Deputy Party Secretary of the TAR branch of the CCP, compliments the monasteries of Drepung and Sera for their implementation of “four standards” education.

Semantic searching can be used along with topic modelling to identify and explore discourses within a corpus. Having located articles with relevant paragraphs makes it easy to go directly to the article. The search program also returns the entire paragraph for each chunk. Metadata is prepended to the displayed text that identifies the original *Tibet Daily* article (an HTML file) in the corpus where the chunk is located and its publication date. A unique number is assigned to every paragraph in the entire corpus. The search program also returns and displays the full paragraph from which the chunk is taken. A measure of similarity to the query text is also shown (the Cohere version 2.0 multilingual embedding model uses the inner dot product

to measure similarity; unlike cosine similarity, this measure of similarity is not normalised and can be larger than 1).

Translations from Tibetan to English of the paragraphs are provided by the Azure multilingual translation API.<sup>5</sup> There are now several translators available for modern Tibetan. We have found that the Azure translator API (version 3.0) currently provides the best translation of modern newspaper Tibetan, though it still makes both syntactical and lexical mistakes.

Table 1: Semantic Search

No.	Chunk	Original Full Paragraph	Similarity
Qry.	323283.01 content_853060.htm 2018-09-12 རང་ལྗོངས་ཀྱིས་“ཚད་གཞི་ལག་བཞེར་བཅིངས་ལུང་གིས། ལྷོན་ཐོན་གྱི་ བཅུན་ཏུར་ཐག་ཉེད་”ཅེས་པའི་སློབ་གསོ་ལག་ལེན་ཉེད་སློབ་སློབ་ཚུན། ས་གནས་ལག་གིས་ཚད་མཐོའི་མཐོང་ཚེན་བྱས་པ་དང་། ལྷོ་ཚན་ལག་ གིས་ལས་ཀ་ཏུར་ཐག་བསྐྱབས་ཤིང་། དགོན་སྡེའི་གྲྭ་བཅུན་གྱིས་སློབ་ འགྲུལ་ལ་གཞོགས་འདེགས་ཏུར་ཐག་དང་། སློབ་སློབ་ལ་རང་འགྲུལ་ རང་ལྗོངས་པ། རང་རྩོགས་རང་སྲོང་ཚོར་འབྲི་བ་བཅས་བྱས་ཏེ་ལྗོངས་ ཡོངས་ཀྱི་གྲྭ་བཅུན་གྱི་བསམ་སློབ་འཛིན་གཅིག་བྱུར་དང་། བོད་ བརྒྱུད་ནང་བསྟན་གྱི་དགོན་སྡེའི་རྒྱན་གཏུན་སློབ་སློབ་སློབ་ཚུན། ཚོས་ ལུགས་ལྷབ་ལོངས་རྒྱན་མཐུན་བཅུན་ལྷོང་དང་། ཡུན་རིང་བརྟན་ལྷོང་། ལྷོན་ཡོངས་བརྟན་ལྷོང་བཅས་འགན་ལེན་བྱུང་ཁར་སློབ་གསོ་ལག་ལེན་ ཉེད་སློབ་སློབ་ལྷབ་ལྷབ་ལྷབ་དུ་བ་བ་ཐོབ་ཡོད། འདི་ག་ཚགས་པར་ཐོག་ དེ་རིང་ནས་བཟུང་“ཚད་གཞི་ལག་བཞེར་བཅིངས་ལུང་གིས། ལྷོན་ཐོན་གྱི་ བཅུན་ཏུར་ཐག་ཉེད་”ཅེས་པའི་ཚད་སློབ་ལེ་ཚན་འདོན་རྒྱ་ཡིན་པས། དོ་ ལུར་ཡོད་པ་ལྟ།	191.53	

<sup>5</sup> <https://azure.microsoft.com/en-us/services/cognitive-services/translator/> (accessed January 15, 2025).

No.	Chunk	Original Full Paragraph	Similarity
1	<p>616415.01 content_121540.html 2021-12-31</p> <p>རང་སྐོར་ལྷོངས་ཏང་ལུད་ཀྱི་རྒྱན་ལུ་འཐབ་ཕྱོགས་གཅིག་ལྷུང་ལུ་ལུ་ལུ་ ཕྱོགས་གཅིག་ལྷུང་ལུ་ལུ་ལུ་ཏང་ཀམ་ཚོང་བརྟན་གྱིས་ཚད་གཞི་བཞེར་བཅུ་སྤྱད་གིས་སྤོན་ཐོན་གྱ་ བཅུན་ཏུར་ཐག་བྱེད་པའི་ལྷོངས་ཡོངས་ཀྱི་སློབ་གསོ་ལག་ལེན་བྱེད་སློ་ སྤེལ་སྤོལ་སློར་གྱི་གནས་ཚུལ་སྤྱད་སེང་ཞུས་པ་རེད། ཚོས་ལུགས་ལས་ རིགས་ཀྱི་འཕུལ་མི་སྤྱབ་ལང་ཐུབ་བསྟན་ས་ལས་གྲུབ་དང་། བཀྲ་ཤིས་ རྒྱལ་མཚན། ཀུན་བཟང་དབང་འདུས། ལླ་བ་ཚེ་རིང་། ལྷ་ཐོག་སློབ་བཟང་ ཡེ་ཤེས། རྗེ་རུང་བསྟན་པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་དབང་ལྷག་སློབ་བཟང་ བསམ་གཏང་བཅས་ཀྱིས་སྤེལ་རེས་གཏམ་བཤད་གནང་ཞིང་། ཚང་ མས་རང་ཉིད་ཀྱི་དོན་དེོས་དང་རྩེ་འབྲེལ་བྱས་ཏེ། སློབ་གསོ་ལག་ ལེན་བྱེད་སློར་ཞུགས་པའི་ཕན་འབྲས་དང་སྤོར་ཚོར་ལུག་ཡོར་སྤོར་ནས། ཚོས་ལུགས་ལས་རིགས་མི་སྲུང་ཏང་ལ་དགའ་འཛིན་དང་། མེས་རྒྱལ་ལ་ དགའ་འཛིན། མི་དམངས་ཀྱི་གཙོ་འཛིན་དང་། སྤྱི་ཚོགས་རིང་ལུགས་ཀྱི་ ལས་ལུགས། མི་རིགས་ས་ཁོངས་རང་སྐོར་ལམ་ལུགས་བཅས་ལ་ མཐའ་གཅིག་ཏུ་བརྟེན་གཏུ། དར་ཆ་གསལ་སྤོན་གྱིས་ལ་ལུ་ལ་ལོ་ ཚོལ། མེས་རྒྱལ་གཅིག་ལྷུང་དང་མི་རིགས་མཐུན་སྦྲེལ་སྤེལ་སྤོར་དང་། ཏང་ གི་བཀའ་ལ་ཉན་པ་དང་། ཏང་གི་བཀའ་རྒྱན་རྒྱ་ཚོར་བ། ཏང་གི་ལྷན་སྐྱེ་ འབྲུང་བ། “མོས་སེམས་ལྷོ་”ཟམ་མི་ཚད་པར་ཟབ་ཏུ་གཏོང་བ། ལྷུང་དུ་ མི་རིགས་གཅིག་མཐུན་འདུས་གྲུབ་ཀྱི་འདུ་ཤེས་བརྟན་པོ་འཛུགས་པ། མོད་བརྒྱུད་ནང་བསྟན་ཀྱང་གོ་ཅན་ལ་སྤེལ་འདེད་གཏོང་བ་བཅས་ཀྱི་ གདེང་ཚིན་དང་ཚིན་སེམས་གང་ལེགས་མཚོན་པར་བྱས་པ་རེད།</p>	<p>616415 content_121540.html 2021-12-31</p> <p>རང་སྐོར་ལྷོངས་ཏང་ལུད་ཀྱི་རྒྱན་ལུ་འཐབ་ཕྱོགས་གཅིག་ལྷུང་ལུ་ལུ་ལུ་ ཏང་ཀམ་ཚོང་བརྟན་གྱིས་ཚད་གཞི་བཞེར་བཅུ་སྤྱད་གིས་སྤོན་ཐོན་གྱ་ བཅུན་ཏུར་ཐག་བྱེད་པའི་ལྷོངས་ཡོངས་ཀྱི་སློབ་གསོ་ལག་ལེན་བྱེད་སློ་ སྤེལ་སྤོལ་སློར་གྱི་གནས་ཚུལ་སྤྱད་སེང་ཞུས་པ་རེད། ཚོས་ལུགས་ལས་ རིགས་ཀྱི་འཕུལ་མི་སྤྱབ་ལང་ཐུབ་བསྟན་ས་ལས་གྲུབ་དང་། བཀྲ་ཤིས་ རྒྱལ་མཚན། ཀུན་བཟང་དབང་འདུས། ལླ་བ་ཚེ་རིང་། ལྷ་ཐོག་སློབ་བཟང་ ཡེ་ཤེས། རྗེ་རུང་བསྟན་པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་དབང་ལྷག་སློབ་བཟང་ བསམ་གཏང་བཅས་ཀྱིས་སྤེལ་རེས་གཏམ་བཤད་གནང་ཞིང་། ཚང་ མས་རང་ཉིད་ཀྱི་དོན་དེོས་དང་རྩེ་འབྲེལ་བྱས་ཏེ། སློབ་གསོ་ལག་ ལེན་བྱེད་སློར་ཞུགས་པའི་ཕན་འབྲས་དང་སྤོར་ཚོར་ལུག་ཡོར་སྤོར་ནས། ཚོས་ལུགས་ལས་རིགས་མི་སྲུང་ཏང་ལ་དགའ་འཛིན་དང་། མེས་རྒྱལ་ལ་ དགའ་འཛིན། མི་དམངས་ཀྱི་གཙོ་འཛིན་དང་། སྤྱི་ཚོགས་རིང་ལུགས་ཀྱི་ ལས་ལུགས། མི་རིགས་ས་ཁོངས་རང་སྐོར་ལམ་ལུགས་བཅས་ལ་ མཐའ་གཅིག་ཏུ་བརྟེན་གཏུ། དར་ཆ་གསལ་སྤོན་གྱིས་ལ་ལུ་ལ་ལོ་ ཚོལ། མེས་རྒྱལ་གཅིག་ལྷུང་དང་མི་རིགས་མཐུན་སྦྲེལ་སྤེལ་སྤོར་དང་། ཏང་ གི་བཀའ་ལ་ཉན་པ་དང་། ཏང་གི་བཀའ་རྒྱན་རྒྱ་ཚོར་བ། ཏང་གི་ལྷན་སྐྱེ་ འབྲུང་བ། “མོས་སེམས་ལྷོ་”ཟམ་མི་ཚད་པར་ཟབ་ཏུ་གཏོང་བ། ལྷུང་དུ་ མི་རིགས་གཅིག་མཐུན་འདུས་གྲུབ་ཀྱི་འདུ་ཤེས་བརྟན་པོ་འཛུགས་པ། མོད་བརྒྱུད་ནང་བསྟན་ཀྱང་གོ་ཅན་ལ་སྤེལ་འདེད་གཏོང་བ་བཅས་ཀྱི་ གདེང་ཚིན་དང་ཚིན་སེམས་གང་ལེགས་མཚོན་པར་བྱས་པ་རེད།</p>	187.09
2	<p>329616.01 content_858688.htm 2018-10-20</p> <p>བསྟན་འཛིན་ནན་བཤད་གནང་དོན། རང་སྐོར་ ལྷོངས་ཏང་ལུད་དང་མིད་གཞུང་གིས་ལྷོངས་ ཡོངས་ཀྱི་ཚོས་ལུགས་ལྷུང་ལོངས་སུ་ཚད་ གཞི་བཞེར་བཅུ་སྤྱད་གིས་སྤོན་ཐོན་གྱ་བཅུན་ཏུར་ ཐག་བྱེད་པའི་སློབ་གསོ་ལག་ལེན་བྱེད་སློ་ སྤེལ་སྤོལ་སློར་གྱི་གནས་ཚུལ་སྤྱད་སེང་ཞུས་ པ་རེད། ཞི་ཅིན་མིང་གི་དུས་རབས་ རབས་གསལ་པའི་ལྷུང་གི་འཕུལ་མི་སྤྱབ་ ཚོགས་རིང་ལུགས་ཀྱི་དགོངས་པ་ཕྱོགས་ ཡོངས་ནས་སློབ་སློར་དང་ལག་བསྟར་གཏོང་ ཐབ་དང་། ཏང་གི་ཚོགས་ཚེན་བཅུ་དགུ་པའི་ དགོངས་དོན་གསལ་པོ་ཤེས་པ་དང་། མཐའ་ ཕྱིན་པ་ཤེས་པ། རོན་འཁྲུལ་ཏན་ཏྱིག་བཅས་ བྱེད་པའི་འཐབ་རྩལ་བཀོད་སྤྱོད་གསལ་ཚེན་ཞིག་ ཡིན་པས།</p>	<p>329616 content_858688.htm 2018-10-20</p> <p>བསྟན་འཛིན་ནན་བཤད་གནང་དོན། རང་སྐོར་ལྷོངས་ཏང་ལུད་དང་མིད་ གཞུང་གིས་ལྷོངས་ཡོངས་ཀྱི་ཚོས་ལུགས་ལྷུང་ལོངས་སུ་ཚད་གཞི་ བཞེར་བཅུ་སྤྱད་གིས་སྤོན་ཐོན་གྱ་བཅུན་ཏུར་ཐག་བྱེད་པའི་སློབ་གསོ་ ལག་ལེན་བྱེད་སློ་སྤེལ་སྤོལ་སློར་གྱི་གནས་ཚུལ་སྤྱད་སེང་ཞུས་ པ་རེད། ཞི་ཅིན་མིང་གི་དུས་རབས་ གསལ་པའི་ལྷུང་གི་འཕུལ་མི་སྤྱབ་ ཚོགས་རིང་ལུགས་ཀྱི་དགོངས་པ་ ཕྱོགས་ཡོངས་ནས་སློབ་སློར་དང་ལག་བསྟར་གཏོང་ ཐབ་དང་། ཏང་གི་ ཚོགས་ཚེན་བཅུ་དགུ་པའི་ དགོངས་དོན་གསལ་པོ་ ཤེས་པ་དང་། མཐའ་ ཕྱིན་པ་ཤེས་པ། རོན་ འཁྲུལ་ཏན་ཏྱིག་བཅས་ བྱེད་པའི་འཐབ་རྩལ་ བཀོད་སྤྱོད་གསལ་ ཚེན་ཞིག་ཡིན་པས། གཞིས་ཚུ་ལྷོང་ལྷོང་ ལྷོ་ཚོན་ལག་གིས་ སློབ་གསོ་ལག་ལེན་ བྱེད་སློ་སྤེལ་བའི་ འཐབ་རྩལ་གྱི་ དོན་སྤོལ་གསལ་ ཚེན་ལ་ལོ་བ་གཏོང་ ཐབ་སྤེལ་ཐོག་གོང་ འོག་མཐུན་ འགུལ་དང་། མཉམ་ ལུགས་ལུགས་གཅིག་ སྤྱོད་ཀྱི་ཚོར་མ་ ཞུགས་པའི་ཏུར་ སེམས་ དང་རང་འགུལ་ རང་བཞིན་འཕེལ་ བར་བྱས་ཏེ། “ཚད་ གཞི་བཞེར་བཅུ་ སྤྱད་གིས་སྤོན་ ཐོན་གྱ་བཅུན་ ཏུར་ཐག་བྱེད་ པའི་སློབ་གསོ་ ལག་ལེན་ བྱེད་སློ་སྤེལ་ སྤོལ་སློར་གཏོང་ ཐབ་ཕྱོགས་སུ་ སྤེལ་དགོས་ ཞེས་ནན་བཤད་ གནང་བ་རེད།</p>	185.76

No.	Chunk	Original Full Paragraph	Similarity
3	329962.01 content_859093.htm 2018-10-23 རྒྱུ་ལོ་ཤར་གྱིས་འབྲས་ལྗངས་དགོན་དང་མེ་ར་དགོན་གྱི་ཚད་གཞི་ བཞི་བརྟེན་སྲུང་ལག་བསྟར་གྱིས་སློན་ཐོན་གྱ་བརྩམ་ཉུང་ཐག་གི་དྲ་ པའི་སློབ་གསོ་ལག་ལེན་བྱེད་སློབ་སྲིད་པའི་ཐད་ཐོབ་པའི་གྲུབ་འབྲས་ལ་ གདེང་འཇོག་གང་ལེགས་གནང་ཞིང་། ཁོང་གིས་དགོན་སྡེ་གཉིས་གྱིས་ ལྟེ་བའི་ལས་རྒྱུད་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་ འགན་ལ་དམ་འཛིན་ནས་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་རབས་གསར་གཏོད་ བྱེད་པ། བྱེད་སློབ་གསོ་ལག་བསྟར་བྱེད་པའི་དང་སྤེལ་བ། མི་དང་། དོན་ པོ། བྱ་བ། བྱ་སློབ་བཅས་མཐོང་ཐུབ་པ་བྱས་ནས་ལྗངས་ཡོངས་ལ་ཆབ་ སྲིད་ཐད་སློབ་འཁོལ་ཚོགས་དང་། ཚོས་ལུགས་ཐད་ཡོན་ཚད་མཐོ་བ། ཀུན་སློབ་ཐད་ཀུན་གྱིས་བཀུར་བ། འགག་ཚུབ་དུས་སྤྱི་ལུགས་པ་ཐོན་ པའི་བཅས་ཀྱི་གྱ་བརྩམ་དུང་འགག་ཅིག་འཇོགས་སློབ་བྱེད་ཚད་སྤེལ་བྱེད་ ཀྱི་ལྗུས་པ་འདོན་སྤེལ་བྱས་ཡོད་ཅེས་གསུངས་པ་ལེན།	185.44	

Table 1: Semantic Search (cont)

No.	Machine-translated Full Paragraph
Qry.	323283 content_853060.htm 2018-09-12 Since our district launched the educational practice of "abiding by the four standards and actively being advanced monks and nuns", all localities have attached great importance to it, and all units should take action. Actively acting, the monks and nuns of the temple actively cooperate with the preaching, take the initiative to participate in learning, consciously write their experiences, and unify the ideological understanding of the monks and nuns in the whole region. Maintain the normal order of Tibetan Buddhist monasteries, ensure sustained and long-term stability in the religious field, and achieve overall stability, and achieve results in educational and practical activities. Starting today, this newspaper will broadcast a column entitled "Abide by the Four Standards and Actively Become Advanced Monks and Nuns." Welcome to pay attention to it.
1	616415 content_121540.html 2021-12-31 Karma Tsedan, Member of the Standing Committee of the Party Committee and Minister of the United Front Work Department of the Autonomous Region, Presents the Educational Practice Activities of the Autonomous Region on "Complying with the Four Standards and Actively Being Advanced Monks and Nuns" Representatives of religious circles Zhukang Thubten Kedrup, Tashi Gangcun, Gongde Wangdui, Dawa Tsering, Mama Lobsang Yeshe, Jalen Tenzin Jebu and Langjie Wangdui. Luosang Jiangcun made an exchange speech, and in the light of their own realities, they talked freely about the results and experiences of participating in the educational practice activities, and guided the religious figures to love the party and the motherland." resolutely support the people's leader, the socialist system and

No.	Machine-translated Full Paragraph
	the system of regional ethnic autonomy, take a clear-cut stand against separatism, and safeguard the unity of the motherland and ethnic unity; It fully embodies the self-confidence and determination to listen to the party, feel the party's kindness, follow the party, continuously enhance the "five identities", firmly establish the sense of community of the Chinese nation, and promote the sinicisation of Tibetan Buddhism.
2	329616 content_858688.htm 2018-10-20 The party committee and government of the autonomous region decided to carry out the educational and practical activities of "abiding by the four standards and actively becoming advanced monks and nuns" in the religious field of the whole region. It is necessary to comprehensively and thoroughly study and implement Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, and understand the spirit of the 19th National Congress of the Communist Party of China. All departments at all levels in Shigatse City should be of great strategic significance to carrying out educational practice activities, and earnestly strengthen the implementation of the strategic deployment of "one belt and one first". It is necessary to deeply understand, link up and down, cohesion, and everyone's participation, so as to stimulate the enthusiasm and initiative of monks and nuns to participate in educational practice activities. We will continue to develop the educational practice of "abiding by the four standards and becoming advanced monks and nuns" in depth.
3	329962 content_859093.htm 2018-10-23 Ding Yexian fully affirmed the achievements of Drepung Monastery and Sera Monastery in the educational practice of "abiding by the four standards and actively being advanced monks and nuns". He stressed that the two sessions closely focused on the central work, vigorously grasped the goals and tasks, innovated work methods, enlivened activities, and adhered to people, things, things, and things. It has played a leading role in building a team of monks and nuns who are "politically reliable, religiously high-level, morally fair, and effective at critical moments" in the whole region.

### 3 Topic Modelling with BERTopic

Topic modelling is a set of techniques used to automatically identify hidden thematic structures within a large collection of documents. It is a form of unsupervised machine learning that does not depend on labels or predefined categories. Over the last two decades LDA (Latent Dirichlet Allocation), first proposed by Blei *et al.* (2003), has been the



most widely used topic model for discovering latent themes in large text corpora.<sup>6</sup> LDA treats each document as a bag-of-words, and by analysing word frequencies, uses a probabilistic generative model to infer which topics are likely represented in each document. However, the availability of transformer models to generate text embeddings has now made possible the use of BERTopic, a state-of-the-art topic modelling tool developed and maintained by Maarten Grootendorst (2022).<sup>7</sup> BERTopic has been used to analyse a variety of corpora assembled from contemporary news sources and political discourse.<sup>8</sup> Topic modelling with BERTopic relies on the same text embeddings as semantic searching. But each text is identified with a single topic (unlike LDA, which treats a document as comprised of a number of topics). Topics can be understood as vector encodings of texts that are semantically similar to each other, and that are clustered together within the high-dimensional semantic space (768 dimensions for Cohere version 2.0).

BERTopic is best described as a pipeline of text processing modules—for text embedding, dimensionality reduction, clustering, and topic representation.<sup>9</sup> It takes transformer-based embeddings, applies a technique like UMAP to reduce dimensionality, clusters documents with a clustering algorithm like HDBSCAN, and then extracts representative keywords to form interpretable topics. Because

---

<sup>6</sup> The Divergent Discourses Project (<https://research.uni-leipzig.de/diverge/>, accessed January 15, 2025) uses the iLCM (integrated Leipzig Corpus Miner) research environment, which implements topic modelling through LDA (<https://ilcm.informatik.uni-leipzig.de>, accessed January 15, 2025; see Kyogoku *et al.* 2025).

<sup>7</sup> BERTopic is available as a Python library with additions and updates on github (<https://github.com/MaartenGr/BERTopic> (accessed January 15, 2025)).

<sup>8</sup> Some examples: Navaretta and Hanson (2023) use BERTopic to generate topic clusters for two policy areas, Energy and Environment, in parliamentary debates and political manifestos by Danish political parties; Aenne *et al.* (2024) use BERTopic to identify dominant themes in two hashtag networks on Instagram, #blacklivesmatter and #blackouttuesday; Xing and Ni (2024) use BERTopic to investigate the portrayal of Maoism in French newspapers in the period 1963-1979.

<sup>9</sup> BERTopic is named for BERT (Bidirectional Encoder Representations from Transformers) and was developed by researchers at Google (Devlin *et al.* 2019). Since 2018 the model has gone through several variations and optimisations, such as RoBERTa, and SBERT, and includes multilingual models as well.

of its modularity it allows for an enormous number of options at every step and comes with a collection of tools for visualising results. By default, BERTopic uses SBERT models from the sentence-transformers library to generate embeddings. However, other embedding models can be used, and in our case we import the Cohere version 2.0 embeddings.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that has some advantages.<sup>10</sup> By examining the local density of points, the algorithm automatically discovers how many clusters (topics) best describe the data, without requiring that the number of topics be specified in advance. It also does not make any assumptions about the shape of the cluster (such as assuming it is spheroid). Topics can be merged, but this is not necessarily desirable if the goal is to characterise a corpus of texts in as much detail as possible. HDBSCAN typically produces a large number of outliers (33% or more). These can be texts that are not recognisably related to any cluster, or they can be texts that fall between clusters. However, this is not a drawback in our case since our goal is not to classify every document but to identify relevant topics. BERTopic also utilises HDBSCAN to extract some “representative paragraphs” whose location within a cluster makes them the best candidates for characterising the cluster.

The clusters discovered by HDBSCAN are mathematical objects. The challenge is to assign meaningful interpretable labels to represent the topics. The default module for BERTopic at this stage in topic modelling is to apply a version of TF-IDF (Term Frequency-Inverse Document Frequency) to all of the documents within each cluster to identify words most uniquely characteristic of each topic compared to the rest of the corpus. The result is five to ten keywords which are then used to label the topic. This necessitates a word tokeniser that splits text into individual words, a list of stop words to be ignored, and a filter for punctuation. The text may also require POS tagging and lemmatisation. This is straightforward for high-resource languages

---

<sup>10</sup> See <https://hdbscan.readthedocs.io/en/latest/index.html> (accessed January 15, 2025).

like English or Chinese, where many such tools exist. But without a reliable word tokeniser for modern Tibetan, it is impossible to create topic representations using TF-IDF.<sup>11</sup>

However, using a generative LLM to read and summarise text and to produce representations of topics is now an option for modern Tibetan. This sidesteps some of the difficulties in developing NLP tools for modern Tibetan. After testing several LLMs, we found that Anthropic's Claude Sonnet 3.5 can read the representative paragraphs for each Tibetan topic directly and generate keywords and labels.<sup>12</sup> The keywords it generates come from the Tibetan source material, eliminating the need for word frequency-based keyword generation.

#### 4 Using BERTopic with the Tibet Daily corpus

Our primary research objective is to see if this form of topic modelling will enable us to identify particular policy or political programs in the TAR, and more specifically to trace changes in their prominence over time. As human readers, we can already recognise from the repetition of certain slogans or keywords in the Chinese or Tibetan media that a policy or drive is underway, but we will not always know when such a drive began or ended. More importantly, the inevitable selectivity of natural reading can mislead us into interpreting a particular policy or drive in terms of its most prominent or striking slogan or keyword – particularly if that keyword indicates unusually repressive measures by a government – and thus overlooking other aspects of that policy or misstating its prevalence in the broader political environment.

Topic modelling with BERTopic can be used with the *Tibet Daily* corpus to offset such tendencies and misreadings. It can uncover the major themes or topics that underly collections of documents,

---

<sup>11</sup> The Divergent Discourses project is developing a Tibetan language model for spaCy that includes a vocabulary tokeniser and POS tagging capabilities to process Tibetan text for input into iCLM. See Kyogoku *et al.* 2025.

<sup>12</sup> <https://www.anthropic.com/claude/sonnet> (accessed January 15, 2025). Claude Sonnet is accessed through an API and accepts detailed instructions through a prompt incorporated into the code.

presenting aspects or purposes of a policy that might not be evident to a casual reader, allowing more precise and comprehensive forms of discourse analysis. The occurrence of topics can be modelled dynamically, revealing when they become prominent and when discussion in the official media tapers off, indicating the life-cycle of each policy or drive. Since topic modelling is based on the same technology as semantic searching, it can identify texts that address a given policy even though none of the keywords normally associated with that policy are used – an important tool for an analyst, since policies and official efforts at ideological promotion in China (and Tibet) are usually put into practice some time before officials settle upon their slogans or keywords. In addition, topic modelling of this type can display the relative prominence or otherwise of a policy within the larger policy environment, revealing themes that might be striking to a casual reader but are perhaps relatively infrequent in overall discourse, or the opposite.

Applying topic modelling to an undefined set of texts, such as the entire contents of a newspaper over a given period, is generally not effective, because the tool will tend to return results that are so general as to be already obvious to the reader, such as “news”, “sports”, “arts” and so forth. We, therefore, selected a particular question of interest and applied the tool only to texts relevant to that question. In this case, we selected “religion” as our overarching question. Our purpose was to see if the tool would provide more insights about the Chinese government’s policies towards religion in the TAR than we had already gathered from unstructured readings of articles over recent years. Those readings had led us to note already a number of key terms or slogans found frequently in official articles and speeches about religion. One example of a religion-related policy term is the phrase “four standards” (ཚད་གཞི་བཞི), discussed earlier in relation to semantic searching, which had seemed to us particularly prominent in our unstructured readings. This term refers to a set of behaviours or attitudes required of all monks and nuns in the TAR. These require the monks to be “politically reliable”, “accomplished in religious knowledge”, “convincing in morality”, and to “play an active role at critical moments” (Chang & Chen 2020). We knew already that these

requirements had been introduced in about 2016 or shortly after (HRW 2018), but we were unsure if the policy would remain in force eight years later. If so, that would be unusual because CCP drives or policies often disappear from public view within two or three years. More importantly, the keywords or formulations (提法 *tifa*) used for each policy or drive are almost never explained in public documents. In the case of the “four standards”, two appear to be about encouraging religious knowledge and ethical conduct, but other religious policies, as we shall see, imply stringent limitations on the actual meanings of these terms. The definition of “an active role at critical moments” could refer to denouncing any others who have dissident opinions or could refer to obeying official orders at the time of the Dalai Lama’s death; it has never been publicly explained. We hoped that topic modelling, combined with semantic searching, would increase our chances of learning more about this drive and its relation to similar policies at the time.

Our first step was to create a subcorpus with a manageable number of articles. To achieve this, we collected all the articles from *Tibet Daily* between 2014 and 2024 in which terms for “religion” appear. We limited ourselves in this study to using the Tibetan-language corpus of *Tibet Daily* rather than the Chinese-language corpus so as to demonstrate how source material in modern Tibetan can be effectively explored using this set of tools.

The articles have been split into numbered paragraphs and then, where necessary, split into chunks of 256 tokens or less. These were filtered paragraph by paragraph for entries that satisfied the Boolean search query: (“བོད་བརྒྱུད་ནང་བསྟན” OR “ཚོས་ལུགས”) — “Tibetan Buddhism” OR “Religion”. Filtering for these two strings (which allows for their appearance in longer phrases) effectively captured all of the paragraphs in which religion was mentioned in some context. The filtered results were then arranged into a CSV file for each year. In total, the religion subcorpus comprised 3,952 articles with 6,622 paragraphs over the 2014–24 period (divided into 9,048 chunks that do not exceed 256 tokens).

In Figure 1 we show a chart, produced using a keyword (string) search on the CSV files in our subcorpus, that displays the relative

frequency of articles (with one or more paragraphs) that satisfy the Boolean search query by year. Articles including one or other of the two terms we used for religion comprise roughly 5% to 7% of all articles between 2014 and 2024. The chart shows, however, a dramatic drop off in the number of such articles from 2022 onwards (as low as 3.5%), an outcome of which we had been completely unaware despite regular reading of the Chinese media in Tibet.

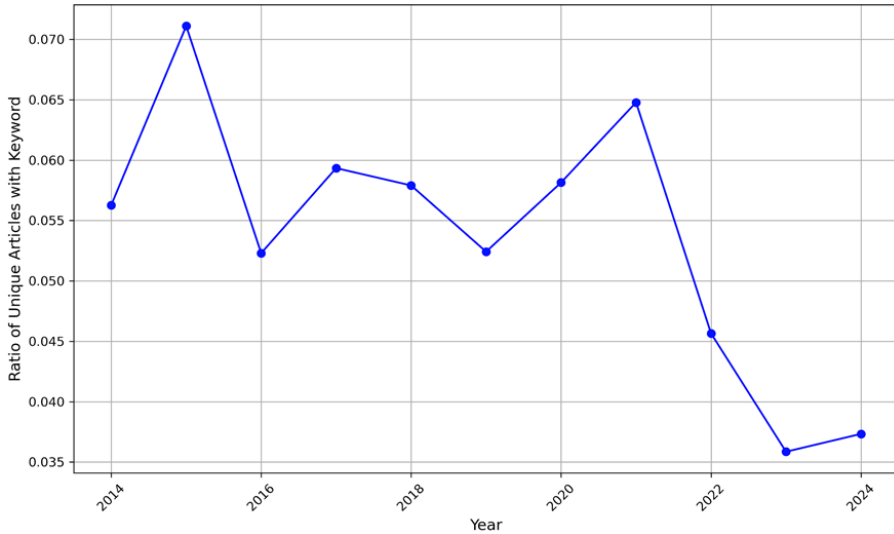


Figure 1 Relative frequency of articles containing the terms བོད་བརྒྱུད་ནང་བཟུང་ (Tibetan Buddhism) or ཚོས་ལུགས་ (religion) in Tibet Daily, 2014–24

We then imported this filtered subcorpus along with the aligned Cohere version 2.0 embeddings for each paragraph chunk into BERTopic. This resulted in 121 topics (0 to 120). There were 4,368 outlier chunks (48.3% of the total of 9,048 chunks) which BERTopic categorised as outliers (it assigns the topic number -1 to them). BERTopic produced what it identified as the ten most representative paragraphs for each topic, and we then used Claude Sonnet 3.5 to read each of these sets of ten paragraphs and to generate topic labels and keywords for them. This produced a label and five keywords in Tibetan, together with English translations, in order of descending importance for each topic. The keywords it selected are not necessarily found in all ten of the representative paragraphs, but taken together

these keywords provide a semantic representation of the ten clustered representative paragraphs for each topic. Table 2 displays a list of the top 21 topics (topics #0 to #20) in descending order of frequency along with the keywords and labels produced by Claude Sonnet 3.5.

Table 2: Topics #0 to #20 with Keywords and Labels

Topic	Count	Keywords	Label
0	256	Religious freedom policy (ཚོས་དང་རང་མོས་སྲིད་ཇུས་), monastery management (དགོན་སྡེ་འདོད་དམ), ethnic unity (མི་རིགས་མཐུན་སྦྲེལ་), legal supervision (ཁྲིམས་ལྟར་དོད་དམ), religious affairs work (ཚོས་ལུགས་ལས་དོན)	Implementation of party's religious policy and monastery management
1	133	Ethnic unity (མི་རིགས་མཐུན་སྦྲེལ་), environmental protection (སྟོན་བཅུད་ལོང་ལུག), religious harmony (ཚོས་ལུགས་ཞི་མཐུན), social stability (སྤྱི་ཚོགས་བརྟན་ལྗིང), people's livelihood (དམངས་འཚོ)	Social harmony and development in Tibet Autonomous Region
2	124	Four standards (ཚད་གཞི་བཞི), model monks and nuns (སྟོན་ཐོན་གྲུ་བཅུན), religious affairs regulations (ཚོས་ལུགས་ལས་དོན་གྱི་སྲོལ་ཡིག), socialist adaptation (སྤྱི་ཚོགས་འཛུགས་དང་འཚམ་མཐུན), Xi Jinping thought (ཞི་ཅིན་ཕིང་གི་དགོངས་པ)	Implementation of four standards policy in Tibetan Buddhist monasteries
3	121	Reincarnation system (སྐུལ་སྐྱེའི་ཡང་སྲིད), government approval (ཡུང་དབྱང་སྲིད་གཞུང་གི་ཚོག་མཚན), legal regulation (ཁྲིམས་སྟོང), traditional procedures (ཚོས་ལུགས་ཀྱི་ཚོག), domestic search requirement (རྒྱལ་ནང་ནས་ཡང་སྲིད་ཚུལ་འཚོལ་བ)	Chinese government control over Tibetan Buddhist reincarnation system
4	113	Religious harmony (ཚོས་ལུགས་འཚམ་མཐུན), monastery administration (དགོན་སྡེ་འདོད་དམ), monk welfare (གྲུ་བཅུན་གྱི་འཚོ་བ), standardised management (ཚད་ལྡན་དོད་དམ), social insurance (འགན་བཅོལ)	Monastery management and religious harmony implementation policies

Topic	Count	Keywords	Label
5	113	United front work (འཐབ་ཕྱོགས་གཅིག་གུང་), religious affairs committee (ཚོས་ལུགས་ལས་དོན་ལྷན་ཁྲུང་), reincarnation training (སྐུལ་སྐྱེའི་གསོ་སྦྱང་), monastery management (དགོན་སྡེའི་དོ་དམས་), Buddhist interpretation (ནང་བསྟན་གྱི་དགོངས་པ་གསར་འགྲེལ་)	Religious and ethnic affairs meetings and training in tibet Autonomous Region
6	104	Patriotic religious devotion (རྒྱལ་གཅིས་ཚོས་གཅིས་), social harmony (སྤྱི་ཚོགས་ཞི་མཐུན་), monastic discipline (སྤྱི་གཞན་ལམ་སྲུང་སྦྱོང་), ethnic unity (མི་རིགས་མཐུན་སྦྲིལ་), Buddhist traditions (བོད་བརྒྱུད་ནང་བསྟན་གྱི་སྲིལ་རྒྱན་)	Buddhist monastics' patriotic and religious development in Tibet
7	100	Social stability (སྤྱི་ཚོགས་བརྟན་ལྷིང་), religious affairs management (ཚོས་ལུགས་ལས་དོན་དོ་དམས་), border security (མཐའ་མཚམས་བདེ་འཇགས་), anti-separatism (ལ་ཕྲལ་ལ་དོ་ཚོལ་), public safety (སྤྱི་ཚོགས་བདེ་འཇགས་)	Social security and religious affairs management in Tibet
8	98	Separatist politics (ལ་ཕྲལ་རིང་ལུགས་), Tibetan independence (བོད་རང་བཙན་), religious exploitation (ཚོས་ལུགས་ཀྱི་ཕྱི་གོས་), social disruption (སྤྱི་ཚོགས་ཟེར་ཟིང་), anti-China forces (གྲུང་གོར་དོ་ཚོལ་)	Chinese government criticism of 14 <sup>th</sup> Dalai Lama's political activities
9	95	Economic development (དཔལ་འཕྱོར་འཕེལ་རྒྱས་), ethnic unity (མི་རིགས་མཐུན་སྦྲིལ་), religious harmony (ཚོས་ལུགས་འཆམ་མཐུན་), social stability (སྤྱི་ཚོགས་བརྟན་ལྷིང་), ecological protection (སྐྱེ་བསམས་སྲུང་སྦྱོང་)	Tibet's modern development and social harmony progress report
10	91	Party gratitude (ཉང་གི་བཀའ་རྒྱུན་), rational religious understanding (ཚོས་ལུགས་ལ་དཔྱད་ཤེས་), poverty alleviation (དབྱུང་སྦྱོང་), happy life (བདེ་སྤྱིད་འཚོ་བ་), ethnic unity (མི་རིགས་མཐུན་སྦྲིལ་)	Party loyalty and religious moderation in economic development
11	86	<i>Thangka</i> paintings (ཐང་གཤམ་), religious artistry (ཚོས་ལུགས་སྐྱུ་རྩལ་), traditional techniques (སྲིལ་རྒྱན་ལག་རྩལ་), monastery displays (དགོན་པར་བཤམས་), artistic periods (དུས་མཚམས་)	Historical development and artistic traditions of Tibetan <i>thangka</i> painting



Topic	Count	Keywords	Label
12	86	Radio translation (རྒྱ་རྒྱུ་ལྷན་ལྷན་སྐྱོང་ལྷན་), cultural adaptation (རིག་གནས་སྲིལ་རེས་), news media (གསར་འགྱུར་བརྒྱུད་ལས་), target audience (གསན་པ་ལོ་), translation accuracy (ཡང་དག་པའི་སྒྲུང་གྲུབ་ས་)	Translation principles and cultural exchange in Tibetan media broadcasting
13	85	Religious criticism (ཚོས་ལུགས་ལ་དཔྱད་ཤེས་), negative influence (གྲགས་ཀྱིན་རན་པ་), present happiness (དེ་ཆའི་བདེ་སྲིད་), show the flag [one's political stance] (དར་ཆ་གསལ་སྟོན་), socialist adaptation (སྤྱི་ཚོགས་རིང་ལུགས་དང་འཛམ་མཐུན་)	Countering religious influence of 14 <sup>th</sup> Dalai Lama through socialist education
14	85	Religious freedom (ཚོས་དད་རང་མོས་), legal religious activities (ཁྲིམས་མཐུན་ཚོས་ལུགས་བྱེད་སྟོན་), counter terrorism (འཛིགས་སྐྱུལ་འིང་ལུགས་ལ་རོ་ཤོལ་), constitutional compliance (བཅའ་ཁྲིམས་སྲུང་བཅི), religious harmony (ཚོས་ལུགས་འཆམ་མཐུན་)	Religious activities regulation and anti-terrorism legal framework
15	83	Religious freedom (ཚོས་དད་རང་མོས་), monastery supervision (དགོན་སྡེ་དོད་ས་), patriotic religion (རྒྱལ་གཅེས་ཚོས་གཅེས་), social harmony (འཆམ་མཐུན་བརྟན་སྲིད་), Buddhist education (ནང་བསྟན་སྲོལ་གྲིང་)	Buddhist monastery management and religious policy implementation in Tibet
16	74	Socialist adaptation (སྤྱི་ཚོགས་རིང་ལུགས་དང་འཛམ་མཐུན་), religious harmony (ཚོས་ལུགས་འཆམ་མཐུན་), monastic management (དགོན་སྡེ་དོད་ས་), Tibetan studies (བོད་རིག་པ་ཞིབ་འཇུག་), academic research (ཐོས་བསམ་ཞིབ་འཇུག་)	Adaptation of Tibetan Buddhism to socialist society and academic development
17	72	Patriotic Buddhism (རྒྱལ་གཅེས་ཚོས་གཅེས་), religious harmony (ཚོས་ལུགས་མཐུན་སྲིལ་), national unity (མེས་རྒྱལ་གཅིག་ཁྱུར་), Buddhist traditions (བོད་བརྒྱུད་ནང་བསྟན་), social development (སྤྱི་ཚོགས་འཕེལ་རྒྱས་)	Panchen Lama's role in Tibetan Buddhism and Chinese socialist society
18	68	Religious patriotism (རྒྱལ་གཅེས་ཚོས་གཅེས་), social harmony (ཞི་མཐུན་སྤྱི་བཀའ་), Buddhist reform (ཚོས་ལུགས་སྒྱུར་བཅོས་), monastic discipline (སྡེ་ཁྲིམས་), social development (སྤྱི་ཚོགས་འཕེལ་རྒྱས་)	Adapting Tibetan Buddhism to modern socialist society

Topic	Count	Keywords	Label
19	64	Rational religious understanding (ཚམས་ལུགས་ལ་དཔྱོད་ཤེས་), happy life (བདེ་སྲིད་འཚོ་བ་), educational guidance (སློབ་གསོ་ཇིང་སྟོན་), hard work and effort (དཀའ་ལྷན་འབད་འཐབ་), reducing religious superstition (ཚམས་ལུགས་ཀྱི་ཕན་མེད་ལུགས་ཀྱི་སེལ་)	Religious education and rational approach to modern life
20	59	Monastery supervision (དགོན་ཕྱེད་དཔལ་), religious harmony (ཚམས་ལུགས་འཆམ་མཐུན་), monastic welfare policies (དགོན་ཕན་གྱ་ཕན་སྲིད་ཅུས་), monks and nuns (གྲུ་བཅུན་), religious development (བསྐྱེད་དོན་ལམ་སྟོར་)	Monastery management and religious policy implementation in Tibet

The information in Table 2 allows us to draw up an overview of religious policy in the TAR during the study period. It shows that “four standards” were indeed an important part of the religious policy environment – they appear as Topic #2 in the clustering performed by BERTopic. In addition, the labels and keywords chosen by Claude Sonnet 3.5 confirm that the four standards are directed at monks and nuns (usually referred to as “religious professionals” in Chinese legal documents) and are part of a regulatory program (“Religious Affairs Regulations”). The inclusion of the word “implementation” in the label given to this topic strongly indicates that the four standards drive involves active engagement by officials in monasteries and nunneries and is intended to require compliance of some sort from its targets immediately (presumably under threat of some kind of institutional punishment, such as expulsion from a monastery). If we look for other instances of implementation-oriented or regulatory drives among these 21 topics, we can see that the labels and keywords for four other topics – #0, #4, #15 and #20 – include the word implementation. This indicates that these topics too primarily concern the management of monasteries (or nunneries) and the imposition of regulations on their personnel. Five other topics (#3, #5, #7, #14 and #16) refer to regulations or management.

Using the labels generated by Claude Sonnet 3.5, we classified all of the 121 topics as belonging to one of four broad categories indicating their political or discursive purpose:

- (1) regulatory topics relating to management or administration;
- (2) celebratory topics praising cultural and social achievements of the state without necessarily declaring a political message;
- (3) ideological-positive topics presenting aspirational concepts, goals and principles; and
- (4) ideological-negative topics attacking or critiquing prevailing beliefs and practices that are to be opposed and eradicated.

In total, regulatory topics comprised 49.9% of paragraph chunks assigned to a topic; celebratory topics represented 8.7% of paragraph chunks; 31.9% belonged to ideological-positive topics; and 9.5% were categorised as ideological-negative topics.

This allows us to hypothesise, if official media pronouncements are indicative, that about one half of religious policy is focused on direct intervention by officials in monastic life and practice or on monastic procedures such as the identification and recognition of reincarnated lamas. These policy drives are thus explicitly directed at strengthening management by the state of “monastic professionals” and their institutions. In terms of print space (or numbers of chunks) dedicated to them in *Tibet Daily*, six of the top eight topics identified by BERTopic involve the implementation of regulations of this kind, primarily on monasteries.

This recalls Pitman Potter’s observation (2003) that religious policy in China in the post-Mao era shifted in the 1990s to a focus primarily on the “management” of “religious personnel” and institutions. Our topic model shows that a focus on the management and regulation of monks and nuns has remained a prominent part of religious policy in the TAR. However, we can also see a significant difference between the current and the earlier regulatory regimes. In the 1990s (specifically from 1996 onwards), the requirements imposed on Tibetan monks and nuns in the TAR required them primarily to supply memorised responses to a written examination, culminating with a declaration of patriotism and a formulaic denunciation of the Dalai Lama (Barnett & Spiegel 1996).

What constituted compliance was thus relatively clear. The four standards (“political reliability”, “accomplishment in religious

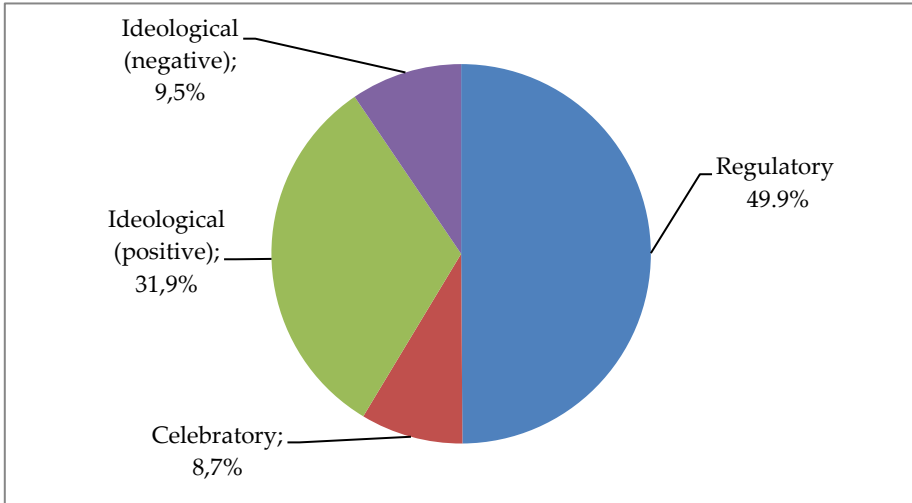


Figure 2 Proportion of paragraph chunks in the Tibet Daily containing regulatory, celebratory or ideological topics relating to religion, 2014–24.

knowledge”, “convincingness in moral conduct”, and being “active at critical moments”) are, however, different in that respect from the 1990s requirements: it is doubtful that they can be legally defined, and compliance would seem hard to measure. There are no signs that formulaic recitation, such as in an examination, is accepted by officials in the current drives as proof of adherence to official demands. Instead, the four standards appear to be a new form of regulatory practice that requires changes to thought, including to thought in the future (“critical moments”), as much as to present behaviour. This reflects the overwhelming shift in Tibet policy, and in policy regarding other “minorities”, under the leadership of Xi Jinping to prioritising the management of thought, not just behaviour (this principle is referred to in some CCP documents as “correctly handling the relationship between ‘controlling the stomach’ and ‘controlling the brain’”; see Chang & Chen 2020).

What, then, does the topic model suggest about the other aspects of religious policy in the region over the last decade? The model does show that some aspects of religious policy are what we might call celebratory, such as promoting “translation accuracy” in media reports (#12), and promoting traditional religious artwork and

painting (#11). These, however, are relatively minor topics, covering only some 8.7% of the text (chunks) in the study corpus. As for the remaining forms of religious policy identified by BERTopic, totalling 41.4% of the paragraph chunks assigned to topics, we can see that they are not regulatory – they do not focus primarily on the imposition of regulations and do not single out monastic institutions as their targets. Instead, they are focused on the promotion of certain concepts, such as “ethnic unity” (མི་རིགས་མཐུན་སྦྲུལ་), “religious harmony” (ཚོས་ལུགས་ཞི་མཐུན་), and “social stability” (སྤྱི་ཚོགས་བརྟན་ལྗིད་). These concepts are not linked in the labels or keywords to regulations or to monastic management, so we can infer that these policies consist of ideological rather than regulatory drives. That is, they are efforts by the state to inculcate, through some sort of educational process, these concepts in the minds of the target group. That group is not primarily “religious professionals”, but the general population. In particular, we can see from the inclusion of such concepts as “people’s livelihood” (རྣམ་པར་འཚོ་), “happy life” (བདེ་སྲིད་འཚོ་བ་), and “hard work and effort” (དཀའ་སྤྱད་འབད་འཐབ་), that these ideological or educational drives are generally aimed at lay believers. If we look closer, we can also see that these drives are not colour-blind – in several cases, such as those about reincarnation, the Dalai Lama or the Panchen Lama, these drives are specifically targeting only Tibetan Buddhists, not followers of other religions. As far as we can tell, these drives may be exhortatory rather than disciplinary – that is, they may not involve explicit punishments or threats (or at least not legal ones).

This emphasis on mass inculcation of certain concepts reflects the instructions given by Xi Jinping at the Seventh Central Forum on Tibet Work in August 2020, where he defined China’s overall Tibet policy for the coming decade:

Xi Jinping pointed out that Tibet work must adhere to the focus and focus on maintaining the unity of the motherland and strengthening ethnic unity. We must strengthen education and guidance for the masses, widely mobilise the masses to participate in the anti-secession struggle, and form an iron wall to maintain stability. We must carry out in-depth education on the history of the Party, the history of New China, the history of reform and opening up, and the history of socialist development, and carry out in-depth education on the history of the

relationship between Tibet and the motherland, and guide the people of all ethnic groups to establish a correct view of the country, history, nation, culture, and religion. We must attach importance to strengthening ideological and political education in schools, run the spirit of patriotism throughout the entire process of education at all levels and types of schools, and plant the seeds of love for China in the hearts of every young person. We must cultivate and practice the core socialist values, and constantly enhance the identification of the people of all ethnic groups with the great motherland, the Chinese nation, Chinese culture, the Communist Party of China, and Socialism with Chinese characteristics. ...We must actively guide Tibetan Buddhism to adapt to socialist society and promote the sinicisation of Tibetan Buddhism. (Xinhuanet 2020)

Before Xi, Tibet policy had generally targeted selected sub-groups of the Tibetan population (primarily monks and nuns, returnees from exile and certain types of intellectuals) as suspected dissidents and subjected them to control and re-education; under Xi, from the Seventh Forum onwards, the primary focus became the “education and guidance” of the Tibetan population as a whole. The topic model shows that a third of the texts in the *Tibet Daily* subcorpus described drives pursuing this new priority.

The labels and keywords generated by Claude Sonnet 3.5 for each topic point to an important distinction among these mass ideological drives: some of them involve positive incentives, while others are negative. Negative propaganda or indoctrination – explicit attacks on or critiques of religious belief – have long been viewed within the CCP as a high-risk strategy when it involves religion. That was the principal reason for Mao’s conciliatory approach to central Tibetans in the 1950s, and for “Document 19”, the famous reformist statement on religion issued by the CPP in 1982, which condemned the anti-religion policies of the Cultural Revolution and called for an end to any attempts by the state to eliminate religion (MacInnis 1989).

Our topic model shows, however, that negative propaganda about religion has become a significant part of the current religious policy environment, although it is confined – at least in print – to a secondary role. Of the texts (chunks) identified by BERTtopic as parts of

ideological drives, 31.9% advance positive goals or concepts, such as “ethnic unity”, “religious harmony”, “social stability”, “patriotic devotion”, “gratitude to the Party”, and “prosperity”. By contrast, the remaining 9.5% appear to signal attacks on certain forms of belief or practice. They include references to “rational religious understanding” (ཚོས་ལུགས་ལ་དཔྱད་ཤེས) and the “negative influence [of religion]” (གུགས་རྒྱན་རན་པ) in topic #13, and to “reducing religious superstition” (ཚོས་ལུགས་ཀྱི་ཕན་མེད་ཤུགས་རྒྱན་སེལ) in topic #19. Similarly, topic #8 appears to be a condemnation of some form of religious view – the label for this topic indicates that this view is related to support for the Dalai Lama – as “religious exploitation” (ཚོས་ལུགས་ཀྱི་ཕྱི་གོས) because it is linked to “separatist politics” (ཁ་ཕྱལ་རིང་ལུགས), “Tibetan independence” (བོད་རང་བཙན), “social disruption” (སྤྱི་ཚོགས་ཟེར་བྱིང), “anti-China forces” (ལྷང་གོང་རོ་རྒྱལ).

In addition, references in the official media to apparently positive concepts such as “rational religious understanding” (ཚོས་ལུགས་ལ་དཔྱད་ཤེས), “happy life” (བདེ་སྦྱིད་འཚོ་བ), “hard work and effort” (དཀའ་སླུང་འབད་འཐབ), and “present happiness” (དེ་ཆའི་བདེ་སྦྱིད) are in practice negative critiques of religion: we know from readings of articles on these topics that “rational religious understanding” and related terms are key parts of critiques of any religious practices deemed excessive, such as offerings to religious figures or institutions. Similarly, “present happiness” is a reference to the Party’s current drive to persuade Tibetans that lay religious belief should never include considerations of one’s future after death. The underlying negative context of the “present happiness” concept is shown by the label for this topic (#13), which describes it as “Countering Religious Influence of 14<sup>th</sup> Dalai Lama”. The topic model thus indicates efforts to present negative critiques of religion in positive terms, but also a failure so far by officials in at least some drives to avoid direct attacks on religious behaviour.

## 5 *Dynamic Topic Modelling*

BERTopic also provides a collection of tools for analysing the evolutions of topics over time through what is known as dynamic topic modelling. Dynamic topic modelling looks at the distribution of the

documents (paragraph chunks) in a topic cluster over a number of timesteps. Spikes in frequency for a topic can signal the occurrence of a political campaign or a new emphasis on propaganda. The keyword representations for topics can be read as the discursive elements that define these developments. Using dynamic topic modelling, BERTopic distributed each of the 121 topics into 11 bins for the period from 1 January 2014 to 14 November 2024. Here we will discuss some of the important topics selected from the first 21 topics (#0 to #20).

Figure 3, for example, which plots topics #13 and #19, allows us to establish a timeline for negative campaigning by officials with regard to religious belief among lay Tibetans. It also shows the discursive interplay of positive and negative language in campaigns. The focus of topic #13 (“Countering Religious Influence of 14th Dalai Lama Through Socialist Education”) is the “negative influence” of an allegedly backward version of Tibetan Buddhism; but it tries to express this critique in a positive way. It does this by stressing the need for a “rational understanding” of religion that will produce happiness in one’s current life. References to this topic or drive first appeared in *Tibet Daily* in 2018 and peaked in 2020 at the time of Xi’s address to the Seventh Forum on Tibet Work. References to this theme fall off quickly by the end of 2022. This suggests that policies are often introduced in a local region some two years or more before a central leader announces them to the public; a meeting such as the Seventh Forum is thus in many ways a confirmation of already existing policies. We see here that an ideological drive or campaign of this sort is relatively short, in this case lasting for around two years, suggesting that they are a response to an immediate but likely short-term instruction from the leadership.

Figure 3 also shows that topic #19 (“Religious Education and Rational Approach to Modern Life”) is coterminous with topic #13. Topic #19 is directed to cadres, instructing them to use a positive form of ideological education regarding religious belief by educating the masses to devote themselves to hard work in this life in order to achieve happiness. Its aim, however, is again a negative one, the “reduction of religious superstition”. These two drives, both active between 2018 and 2022, illustrate the use of both positive and negative



discursive forms in propaganda, but it is clear that the anti-Dalai Lama drive at this time was significantly more prominent than the drive to persuade believers to focus exclusively on their “present happiness”.

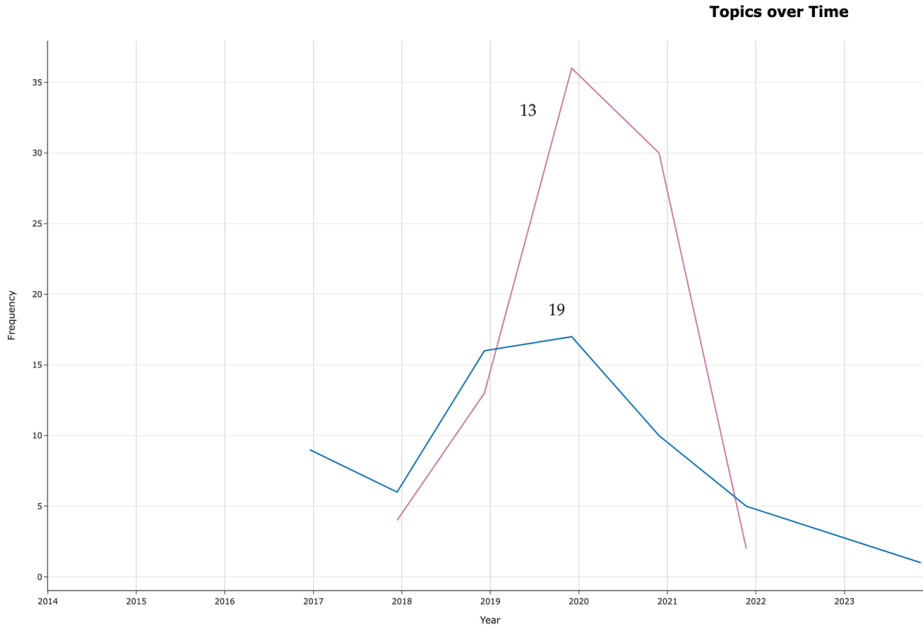


Figure 3 Frequency in the Tibet Daily of topics #13 (“Countering the 14th Dalai Lama”) and #19 (“Religious Education and Rational Approach to Modern Life”), 2014–24.

Figure 4 provides timelines for three drives that had somewhat longer lifespans. These were either regulatory drives or were fundamental to China’s long-term Tibet policy. The first and most prominent is the drive to impose governmental regulations concerning the selection of reincarnate lamas (topic #3). This became a major political priority for the Chinese state after the exiled Dalai Lama unilaterally declared his recognition of a child as the 11<sup>th</sup> Panchen Lama in May 1995; this led China to impose formal regulations in 2007 abrogating to itself alone the right to choose or appoint reincarnate lamas. The drive to enforce compliance with these regulations accelerated after 2011, when China appears to have begun preparing for the death of the current Dalai Lama. We see accordingly that this drive was already in process at the start of the period covered by our subcorpus in 2014, that it peaked in 2019 and again to a lesser extent in 2021, and still continues. The long-

running nature of this drive is as expected, but the peak in 2019 has yet to be explained, and, once again, the marked decline in visibility of this drive (and of any *Tibet Daily* articles referring to religion) after 2022 is surprising.

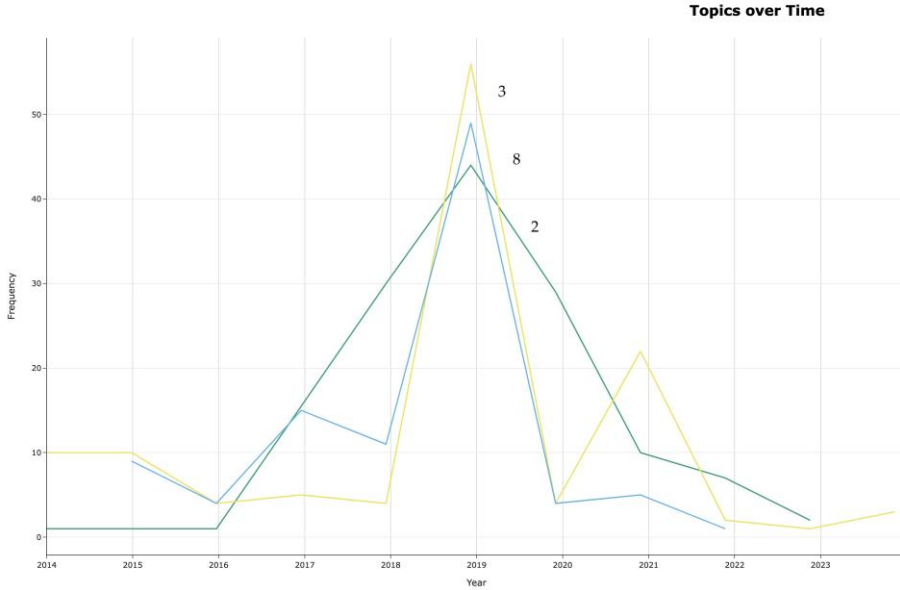


Figure 4 Frequency in the *Tibet Daily* of topic #3 (“Government Control Over Tibetan Buddhist Reincarnation System”); topic #8 (“Criticism of the 14<sup>th</sup> Dalai Lama’s Political Activities”); and topic #2 (“Implementation of Four Standards Policy in Tibetan Buddhism”), 2014–24.

The second timeline shown in Figure 4 maps a closely related topic, the drive to denounce “the 14<sup>th</sup> Dalai Lama’s Political Activities” (topic #8). This, the most striking and controversial of all Chinese policies in Tibet, was initiated at the Third National Forum on Tibet Work in July 1994, is generally seen as in many ways the bedrock of China’s political strategy in Tibet. We see here that it continued into the 2020s. Surprisingly, however, it seems to have been out of sight in 2014, to have re-emerged in 2015, and to have peaked in 2019 shortly before the Seventh Forum, exactly at the same time as the drive to promote the reincarnation regulations. Denunciations of the Dalai Lama (or “the Dalai”, as he is referred to in the official Chinese media) in *Tibet Daily*, at least in the form identified by BERTopic here, then

disappeared from view shortly afterwards, with no references since 2022. These again are findings that have not previously been noted.

The third topic shown in this figure is the four standards drive (topic #2), which, as we have seen, is a regulatory drive that imposed new requirements on monks and nuns. The timeline shows that it was already appearing as a topic in *Tibet Daily* by 2014, at least two years before the time when foreign observers had believed it to have begun. References to this drive do not appear after 2023. Though these three topics are directed at different targets (the “four standards” campaign in topic #2 is for the monasteries, while the anti-Dalai Lama drive and the promotion of reincarnation regulations are society-wide), they overlap very precisely, all peaking at the same time in 2019. We can see that the recent peak in these regulatory or long-running drives (the anti-Dalai Lama and the reincarnation drives would certainly have shown earlier peaks in the period before 2014, if our subcorpus had included those years) occurred a year before the Seventh Forum and the corresponding peak in the ideological drives that we saw in Figure 3. Overall, we can see that drives run in tandem: at a time when the CCP activates a push on religion in the TAR, that push will consist of multiple components and subsidiary drives more or less simultaneously.

Dynamic topic modelling also shows topics which are not marked by a single peak or a short duration, but are long-standing and recurring. These are distributed more evenly over a number of years and indicate a discourse that is sustained over a longer period or repeated regularly. Some examples are displayed in Figure 5; all these topics or drives continued throughout our research period and were still being referred to in *Tibet Daily* as of 2024. The most prominent concerns monastery management (topic #0). This is the highest frequency topic identified by BERTopic.. It is a regulatory drive directed at monasteries, which, while restating the principle of respecting religious beliefs and allowing normal religious activities, cracks down on illegal activities and calls on cadres to implement and strengthen the Party’s management of religious institutions. It is present throughout the entire period from 2014 to 2024, rising from 2015 onwards, falling off to some extent by 2020, with a sustained peak

between 2016 and 2018, and again in 2022. It exemplifies the principle that we have already seen that religious policy in Tibet is at its basis the persistent imposition of managerial control over monks and nuns. When the monastery management drive drops off slightly in 2019, we can see that it is replaced by the peak in the four standards drive (topic #2), which is only a new and more demanding form of monastic management, as well as by the drives requiring compliance with the anti-Dalai Lama drive and the reincarnation regulations.

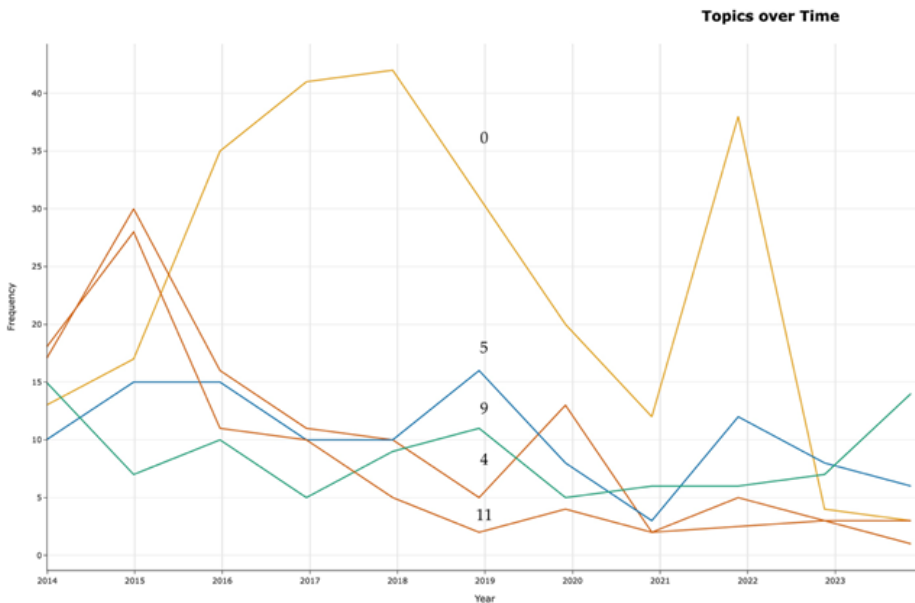


Figure 5 Frequency in the Tibet Daily of topic #0 (“Implementation of Party’s Religious Policy and Monastery Management”); topic #4 (“Monastery Management and Religious Harmony Implementation Policies”); topic #11 (“Historical Development and Artistic Traditions of Tibetan Thangka Painting”); topic #5 (“Religious and Ethnic Affairs Meetings and Training in Tibet Autonomous Region”); and topic #9 (“Tibet’s Modern Development and Social Harmony Progress Report”), 2014–24.

The timelines in Figure 5 show a number of other topics or drives that are long-running and persistent, marked by continuity rather than peaks. These topics include two subsidiary components of the monastery management drive that evidently are more or less in continuous operation. One – “Monastery Management and Religious Harmony Implementation Policies” (topic #4) – is basically a reminder to cadres to present, as its label indicates, a positive aspect of monastic

regulation by emphasising the state's provisions of health and welfare support for (officially recognised) monks and nuns. The other is the ongoing training and oversight of the cadres who implement religious affairs policies in Tibet, under the leadership of the Party agency known as the United Front Work Department (topic #5).

Other "persistent" topics shown in Figure 5 are one celebrating Tibetan religious painting (topic #11), and one linking "religious harmony" to economic development (topic #9). These topics are more likely to represent consistent themes to be maintained in propaganda work rather than drives. Here again we see an effort to emphasise positive forms of propaganda and management.

## 6 Conclusion

This discussion of topic modelling has focused on just a few of the higher frequency topics that were identified in our research corpus. But even from this selection, we can reach a number of provisional hypotheses. One is that high-frequency topics in some cases will represent political drives or campaigns where officials are mobilised to achieve a specific outcome in one or other sector of society. Such drives will typically be of relatively short duration, perhaps of one to two years in some cases, and will be marked by peaks in terms of frequency of references in the media. These will often be responses to instructions or calls from a central leader, and may be intended to signal highly visible compliance by local officials to national-level instructions.

In general, we noted two kinds of drives of that type: regulatory ones aimed mainly at religious institutions and professionals, and ideological ones that aim to change thought and attitude among the wider public. The regulatory drives, and above all monastery management, appear to be the basic, ongoing or staple element of religious policy in the TAR. The regulatory drives will often be of relatively long duration, and will include multiple subsidiary drives.

The topics that indicate ideological drives, designed to inculcate a particular concept or opinion among the population, will typically be

shorter in duration and stronger in intensity. These drives appear to be a dominant feature in the Xi Jinping era, since the entirety of minority populations are now deemed in need of radical political re-education and improvement. They show recurrent attempts to present negative critiques of religion in positive terms.

Topics which involve denunciations of the Dalai Lama or of Tibetan independence (and required compliance with reincarnation regulations) are an exception to the principle of avoiding negativity in religious discourses. These topics show peaks of activity but are persistent over time. This appears to reflect the persistent perception among officials of the Dalai Lama and the independence concept as the core threat to China in Tibet, the core assumption on which all Tibet policy has depended since the mid-1990s. In general, specific discourses that attack the Dalai Lama or the concept of independence precede ideological drives that critique non-approved forms of religious behaviour; the former are more likely to persist.

We also identified a minor type of topic that consists of celebratory discourses that emphasise the state's support for the positive role of (reformed or improved) religion in the economy, art or media. These topics are relatively low-frequency but persistent, suggesting that they mark themes in propaganda or rhetoric rather than specific drives.

Overall, the topic model showed that, besides topics that indicate time-specific drives and ongoing core political themes, there are also a large number of topics that we call "maintenance themes" — concepts, arguments, opinions and insistences that are low-frequency, but persist throughout the research period. They provide a kind of sustained intellectual continuo to the peaks and troughs of regulatory and ideological drives. The anti-Dalai and anti-splittist discourses are also of this type, but are far more prominent and virulent in their profile than most such background themes.

To fully understand and interpret a topic, it is necessary to delve into the representative paragraphs that BERTopic assigns to each cluster. Our objective here has been to demonstrate how topic modelling with BERTopic, employing transformer-based numerical encoding of text, can be used effectively to investigate a corpus of Tibetan documents like *Tibet Daily*. By leveraging modern LLMs (such

as the one from Cohere), which support embeddings for Tibetan texts, we can bypass some of the challenges associated with traditional topic modelling methods like LDA. Given the nature of *Tibet Daily* as a source, with its stated role as a government or Party organ, we have interpreted the topics identified by our model as indicators of political campaigns, ideological drives, the dissemination of policy, and the ongoing shaping of public opinion on major issues relating to religion. Nevertheless, our interpretations are limited, particularly because *Tibet Daily* is only one of many official outlets used by the Tibet authorities, and because it reflects provincial-level priorities, not those at a local level. When a political campaign, new policy directive, or organisational imperative for Party and government cadres disappears from sight in the columns of *Tibet Daily*, it may signal not that that campaign or policy has ended, but that it is at that time being implemented throughout the TAR at the local level. In principle, we would therefore aim to expand the sources for our topic model to include local as well as provincial-level media in Tibet, a difficult project. Nevertheless, the use of topic modelling on media at the provincial level provides an abundance of information and insights about the complex, multiform nature of policy implementation and propaganda practices in the TAR.

### Bibliography

Barnett, Robert, and Mickey Spiegel.

*Cutting Off the Serpent's Head: Tightening Control in Tibet, 1994–1995*. London: Tibet Information Network and New York: Human Rights Watch, 1996.

Blei, David M., Andrew Y. Ng, Michael I. Jordan

“Latent Dirichlet Allocation,” *Journal of Machine Learning Research* vol. 3, 2003, pp. 993-1022. Available online at <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> (accessed January 26, 2025).

Chang Chuan 常川 and Chen Yuejun 陈跃军

“吴英杰：狠抓既定维稳措施落实 确保社会大局和谐稳定” (Wu Yingjie: Pay Close Attention to the Implementation of the Established Stability Maintenance Measures to Ensure the Harmony and Stability of the Overall Social Situation). 西藏日报 (*Tibet Daily*). Posted on 共产党新闻网 [Chinese Communist Party News Network], 26 August, 2020. <http://cpc.people.com.cn/n1/2020/0826/c64102-31837539.html>.

Chang Chuan 常川, Chen Zhiqiang 陈志强 and Chen Yuejun 陈跃军

“西藏自治区代表团向昌都解放纪念碑敬献花篮、看望驻昌 部队官兵、宗教界人士并与各族各界代表座谈 吴英杰讲话” [The Delegation of the Tibet Autonomous Region Presented Flower Baskets to the Chamdo Liberation Monument, Visited the Officers and Soldiers of the Troops Stationed in Changdu, Religious Figures, and Held Discussions with Representatives of All Ethnic Groups and Walks of Life: Wu Yingjie Delivered a Speech]. 西藏日报 (*Tibet Daily*), reposted by cpcnews.cn, 11 October, 2020. Available online at <http://cpc.people.com.cn/n1/2020/1011/c117005-31887459.html> (accessed January 20, 2025).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.

“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *North American Chapter of the Association for Computational Linguistics*, 2019. Available online at <https://arxiv.org/pdf/1810.04805> (accessed January 15, 2025).

Engels, James, and Robert Barnett

“Developing a Semantic Search Engine for Modern Tibetan”, *Revue d'Etudes Tibétaines* 74, 2025, pp. 262–283.

Goldstein, Melvyn C., Ben Jiao, and Tanzen Lhundrup.

*On the Cultural Revolution in Tibet: The Nyemo Incident of 1969*. Berkeley, CA: University of California Press, 2009.

Grootendorst, Maarten

“BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure,” *arXiv preprint*, 2022. [doi:10.48550/arXiv.2203.05794](https://doi.org/10.48550/arXiv.2203.05794).



Hartley, Lauran

“Tibetan Publishing in the Early Post-Mao Period,” *Cahiers d’Extrême-Asie* (15), pp. 231-252. [doi:10.3406/asiae.2005.1227](https://doi.org/10.3406/asiae.2005.1227)

HRW (Human Rights Watch)

“China: New Political Requirements for Tibetan Monastics. Authorities ‘Sinicizing’ Religion,” *Human Rights Watch*, 2018. Available online at <https://www.hrw.org/news/2018/10/30/china-new-political-requirements-tibetan-monastics> (accessed January 15, 2025).

Knierim, Aenne, Michael Achmann, Ulrich Heid and Christian Wolf

“Divergent Discourses: A Comparative Examination of Blackout Tuesday and #BlackLivesMatter on Instagram,” *CLiC-it 2024: Tenth Italian Conference on Computational Linguistics*, 2024. Available online at [https://ceur-ws.org/Vol-3878/53\\_main\\_long.pdf](https://ceur-ws.org/Vol-3878/53_main_long.pdf) (accessed January 26, 2025).

Kyogoku, Yuki, Franz Xaver Erhard, James Engels, and Robert Barnett

“LLM in Low-resourced language NLP: Developing a Basic spaCy Modern Tibetan Language Model from Scratch,” *Revue d’Etudes Tibétaines* 74, 2025, pp. 187–220.

MacInnis, Donald E.

*Religion in China Today: Policy and Practice*. Maryknoll NY: Orbis, 1989.

Meelen, Marieke

“Classical Tibetan Word Embeddings (Version 1),” [Data set]. *Zenodo*, 2022. [doi:10.5281/zenodo.6782247](https://doi.org/10.5281/zenodo.6782247).

Navarretta, Costanza and Hansen, Dorte Haltrup

“According to BERTopic, what do Danish Parties Debate on when they Address Energy and Environment?” In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, Association for Computational Linguistics, 2023, pp. 59–

68. Available online at <https://aclanthology.org/2023.cpss-1.6.pdf> (accessed January 26, 2025).

Potter, Pitman B.

“Belief in Control: Regulation of Religion in China,” *The China Quarterly* 174, 2003, pp. 317–337. [doi:10.1017/S0009443903000202](https://doi.org/10.1017/S0009443903000202).

Sabbagh, Christina

“Improving alignment for low-resource parallel corpora”, Master of Science, Speech and Language Processing, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 2023.

Tsering Wooser.

*Forbidden Memory: Tibet during the Cultural Revolution*. Lincoln, NA: Potomac Press, 2020.

Xinhuanet

“习近平：全面贯彻新时代党的治藏方略 建设团结富裕文明和谐美丽的社会主义现代化新西藏” [Xi Jinping: Comprehensively Implement the Party’s Strategy for Governing Tibet in the New Era and Build a Socialist Modernised Tibet That is United, Prosperous, Civilised, Harmonious and Beautiful]. *Xinhuanet*, 29 August, 2020. Available online at [http://www.xinhuanet.com/politics/leaders/2020-08/29/c\\_1126428221.htm](http://www.xinhuanet.com/politics/leaders/2020-08/29/c_1126428221.htm) (accessed January 20, 2025).

Xing, Ying and Wenjing Ni

““Mao Fever” in France: The Reception of Maoism in the French Mass Media, 1963-1979,” *American Journal of Chinese Studies* 31 (1), 2024, pp. 1-24.

Zhang Xiaoming.

*China’s Tibet*. China Intercontinental Press, 2004.

Zhao Shenying

西藏风云 [Tibet Storm]. Beijing: Xinhua Publishing House, 1987. Available online at <https://books.google.com/books?id=-IPTAAAAMAAJ> (accessed January 25, 2025).

*Appendix: A global display of all of the topics*

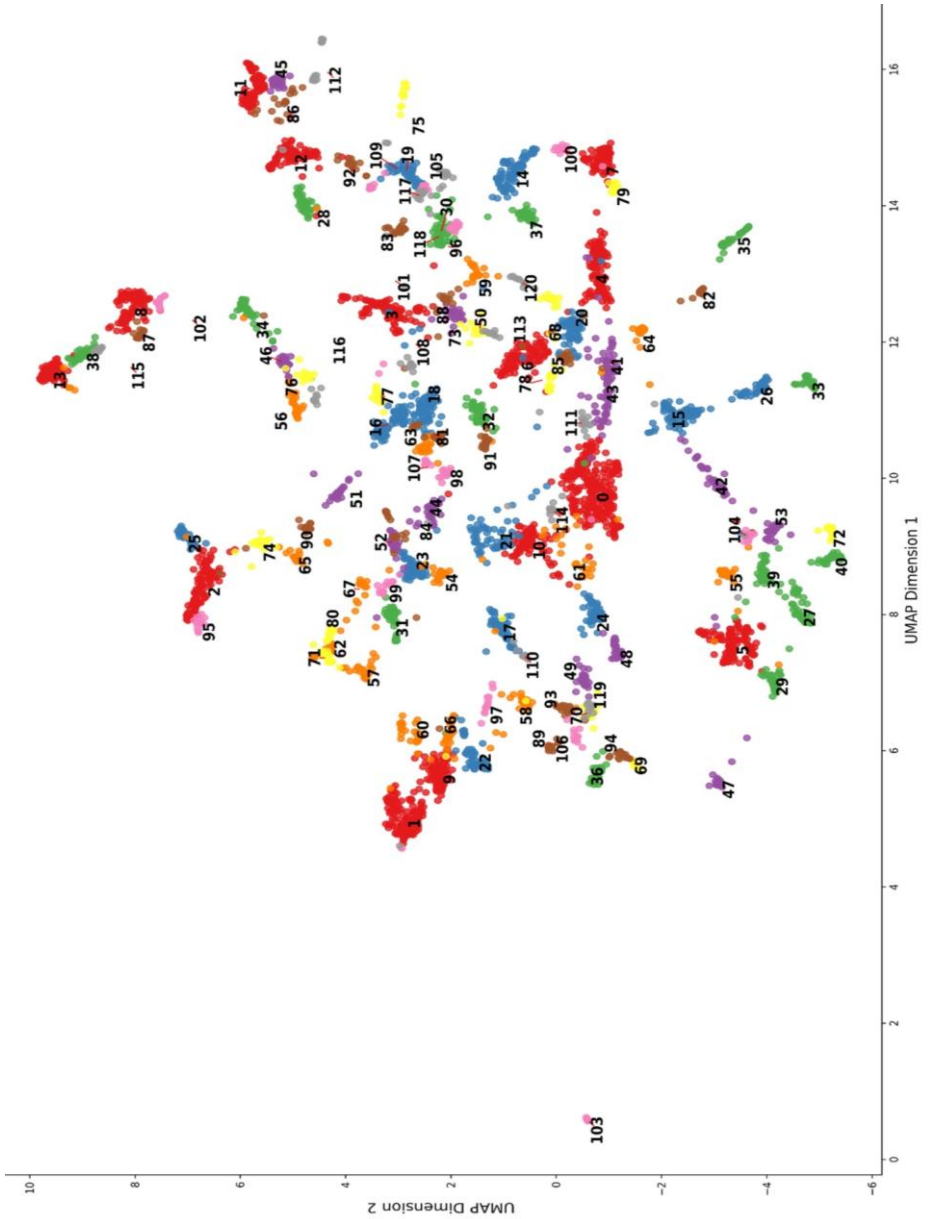


Figure 6 UMAP Projection in 2 dimensions of 121 Topic Clusters Extracted Using BERTopic

The UMAP plot in Figure 6 displays the distribution of all clusters within the latent semantic space. The 121 topics (0-120) are each coloured and numbered. The UMAP algorithm reduces the dimensionality of the 768-dimension vector space to two dimensions for visualisation purposes, ensuring that points that are close in the high-dimensional space remain near each other in the two-dimensional visualisation.<sup>13</sup>




---

<sup>13</sup> Topic #103 reports speeches in *Tibet Daily* by Li Kexiang over the entire period. These are not about Tibet, but mention religion (རྗེས་ལྷན་ལས), and thus were included in the subcorpus. In Figure 6 the cluster for Topic #103 is far to the left.

# Developing a Semantic Search Engine for Modern Tibetan

James Engels (University of Edinburgh)  
and  
Robert Barnett (SOAS University of London)\*

atural language processing (NLP) as a scientific and mathematical realm has undergone at least two generational shifts since the first earnest work on Tibetan NLP began in the late 1990s and early 2000s (e.g., Hackett 2000, Jiang 2003). The first generation of generic NLP tools was developed from theoretical foundations that were established as early as the 1950s – implementations in computers of (theoretically) crosslinguistically applicable formalisms that could be used to create relatively small-scale tokenisers and parsers. The practical realisation of the second and third generations of NLP development arrived soon after the publication of the first mathematical proposals on which they rely.<sup>1</sup>

---

\* The research for this paper was conducted as part of the Divergent Discourses project which received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (<https://gepris.dfg.de/gepris/projekt/508232945?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/>. Coding for all tasks described in this paper was carried out by James Engels.

<sup>1</sup> Blei *et al.* (2003) provide a reasonable benchmark for the original mathematics of generation II, in which the authors proposed Latent Dirichlet Allocation (LDA), a method still used for topic modelling across small-scale applications. The transformer revolution characterises Generation III, starting with the landmark paper “Attention is All You Need” by Vaswani *et al.* (2017). The impact of “Attention” was immediate and overwhelming.

Despite consistent effort by a small number of dedicated researchers, the availability of tools from the most recent (post-transformer) generation of NLP advancement depended on parseable corpora at a scale that was not achievable in Tibetan, at least until roughly 2020. Recent work by Marieke Meelen and collaborators has contributed greatly to language-technological support for classical Tibetan, either in the form of preliminary parsing models and treebanks (Faggionato & Meelen 2019) or large parsed corpora (Meelen *et al.* 2021).

This article is concerned with the development, for the first time, of a semantic search engine for modern Tibetan. Semantic search contrasts with keyword search in its capacity to compare similarity of meaning across queries and results, independent of literal lexical overlap. To illustrate this capacity, we briefly describe how this is achieved, first by contrasting it with a prototypical keyword search method and then by discussing the major components of a semantic search system that is under development for general use.

Keyword searches were the standard for search systems for most of the life of the internet – indeed, when most people imagine a search system, they usually imagine a keyword search. To make keyword searches useable, the system must reward query results that contain “useful” words – that is, keywords that can be used to diagnose a more relevant result – and punish more common words by, essentially, ignoring them. To some extent, this is achieved through lists of stopwords – very common words, such as “is” or “the”, which are usually pre-excluded from searches. Thus, the query “average weight of an elephant” would be reduced to “average weight elephant” for maximum relevance, and this can be done by systematically removing all instances of “of” and “an.” But for more complex cases, where much less frequent terms mix with much more frequent terms, such as a query about a “new building site overseen by the Ministry of Finance”, it is the combination of relatively frequent terms that generates the most relevant results. Most keyword search systems typically guess at the relative importance of words in a query by checking for their frequency in the overall corpus. Even simpler

versions of keyword searches check whether strings in their searchable data contain the same strings as in the query.

Semantic search, on the other hand, allows the user to enter a long query (ideally two to three sentences or the equivalent) and to receive results based on the similarity of the content of the returned result, independent of keyword similarity. Semantic search systems inherently require a large corpus from which to build a semantic space or a prefabricated semantic space that can be secondarily imported and used to generate the search space for the smaller corpus. This contrasts with most keyword search systems. To illustrate, consider our previous “Ministry of Finance building site” example: a human reader with world-knowledge might be able to deduce that a different text about a “newly constructed tax office” in the same region is likely the same building, but with no overlapping keywords, that text would be invisible to a standard keyword search. A well-trained semantic search system will identify the underlying conceptual overlap between the two queries and will return results containing the second to a user querying the first.

For well-resourced languages with well-developed lower-level NLP tools, creating a facsimile of that language’s semantic space is computationally intensive but resource-trivial because the training data and/or the frameworks are already readily available and widely known. The most important component of such a process is a reliable tokeniser – a tool that can reliably identify and split words. In white-spaced languages like those found in Europe, this begins with the relatively trivial step of splitting text on white spaces and punctuation and compiling them into a list. Languages of East and Southeast Asia, on the other hand, tend to identify syllables using overt marking but do not differentiate in writing between, say, two consecutive monosyllabic words and a disyllabic word (for more on the tokenisation problem in Tibetan, see Meelen *et al.* 2021 and Kyogoku *et al.* 2025 in this issue).

What is to be done? Small language models rely on rule-based tokenizers. In languages without word spacing, this creates a significant challenge. These models must use long and complex lookup algorithms. The process works by checking syllables one by

one against a large hidden dictionary. If no match is found, the algorithm checks pairs of syllables for possible entries. This continues until a likely match is identified. The process then repeats with any remaining syllables until the entire text is processed.

Large language models (LLMs) like GPT work quite differently. They function more like human readers. This is achieved through their training process. They are trained on billions of texts and trillions of words in natural contexts. Through this exposure, the models develop an understanding of word distribution patterns. As a result, they develop something akin to natural language understanding. This allows them to parse text in a way that resembles how proficient human readers process language.

The ability of an LLM to tokenise is a secondary property of what might be called its “understanding space,” an abstract geometry of relationships between all its word-concepts. An Understanding Space can be imagined as a complex map showing how all words and concepts relate to each other. This map is created through training on massive amounts of text data. While it cannot be perfect (since that would require learning from every possible sentence), it gets remarkably close to human-like understanding, generated by training on billions of examples. The relationships between words, and indeed entire texts/utterances, is a literal projection onto the multidimensional space that the machine generates from its past training, and once the training is complete, that understanding space becomes fixed.

Once this understanding map is complete, it's used to create a second system, a second multidimensional space (let's call it the “Memory Space”), into which texts can be projected, i.e., where actual documents get stored for searching. When the semantic search database is first generated, texts are converted into mathematical representations or vectors by projecting the text content into the memory space using the understanding space.

Once the Understanding Space is set, new search results can be indefinitely added to the Memory Space using the same Understanding Space without the need to retrain anything. When the search tool actually conducts a search, the user's query is projected



temporarily into the memory space and compared to its most similar stored vectors corresponding to documents.

Training an understanding space from scratch for a low-resource language is no easy task, requiring gigabytes to terabytes of well-organised text. We found, however, that Cohere, a company founded by former Google scientists, had released access to understanding spaces for many low-resource languages to little fanfare. Its default embedding model includes native support for Tibetan, and its Multilingual Model 2.0 provides an out-of-the-box understanding space for Tibetan.<sup>2</sup> It also claims to provide support for, among other languages, Arabic-script Uyghur and (Cyrillic) Mongolian.

More recently, other large corporate entities have incorporated Tibetan into their recent wave of language technology offerings, with varying degrees of success. Google Translate introduced Tibetan in June of 2024, though proficient Tibetan readers are generally unsatisfied with its capabilities, impressionistically comparing it to Google Translate's powers in Spanish and French from its earliest days. OpenAI significantly improved its Tibetan proficiency between GPT-4 (2023) and GPT-4o (2024), generating passable translations from English to Tibetan and back again. In general, by late 2024, Tibetan language understanding and text generation from corporate AI services had progressed from being practically useless curiosities to reliable everyday tools for Tibetan text analysis. However, while these tools are useful for a single user with a single query at, say, the ChatGPT interface, their capacity for use at larger scales requires a level of programming knowledge that not all Tibetologists have, to say nothing of the high upfront costs for constant API requests.

Given this evolving context in the new capabilities available for machine translation and NLP of Tibetan, the rest of this paper is intended to achieve two goals. The first is to familiarise less technically proficient Tibetologists with the broad theories required to build such a system, in roughly chronological order of their appearance and

---

<sup>2</sup> Cohere's user policies are additionally attractive because the vectors that their model generates are the property of the user who generated them.

application in NLP systems. The second is to explain the structure of our semantic search system, which relies in differing amounts on theoretical elements from each of the sections.

## 1 *Vectorisation*

Our project, *Divergent Discourses*, includes two main channels for developing computational tools for use with modern Tibetan. One involves adapting an existing tool, the integrated Leipzig Corpus Miner (iLCM), to carry out complex forms of keyword searches in modern Tibetan using an approach based on a method called Latent Dirichlet Allocation (LDA, developed by Blei *et al.* 2003; see below) and an engine provided by the language modelling software package spaCy (see Kyogoku *et al.* 2025 in this issue). The second involved the building of two applications for use with modern Tibetan texts that use a vector-based approach: a topic modelling engine (automatic, corpus-scale identification of textual foci; see Schwartz & Barnett 2025 in this issue) and a semantic search engine.

For the first task, a good tokeniser is the single most important low-level NLP requirement: without it, the process faces a bottleneck. This is also the case with developing Named Entity Recognition (NER) for any language, or if one is creating a language model for a platform such as spaCy. However, vector-based topic modelling and semantic searching do not need tokenising or part-of-speech (POS) tagging to be carried out on their training data.

### 1.1 *Document Vectorisation*

In NLP, a vector is a numerical representation of a word, phrase, sentence, or document that captures its meaning in a way that computers can process. Vectorisation can be imagined as converting a language into a list of numbers that preserve the semantic relationships of its constituencies. One of the earliest methods of vectorisation (by the standards of AI development) is called “term

frequency-inverse document frequency" (TF-IDF), originally proposed by Karen Spärck Jones (1972). In this system, which is still frequently used, including by the iLCM, each document (in NLP, the term "document" is generally used for a paragraph) in a corpus is vectorised, as well as each word. A matrix is created in which the rows consist of the numerical strings ascribed to each document (paragraph) and columns hold the strings for each word in that document. This produces a frequency table, which can be used as a baseline or master table for comparison with matrices produced for other documents, as well as with the weight of terms reflecting their overall frequency in the entire corpus. This by default punishes very common content words and rewards more specific or unusual terms. This generates a matrix from which specific features of documents are easier to extract solely because their numerical values will be further from the mean. This is particularly useful for topic modelling.

Techniques like TF-IDF are usually termed document vectorisation because their dimensionality (size) is exactly the number of unique features across the entire corpus – i.e., the total number of unique non-stopword words. Because each document matrix consists only of the co-occurrences of entries in the lexicon, they are known as *sparse* matrices, where any column will only have one cell containing a value, and all others will be empty or 0. Suppose there are  $D$  unique words in a document and  $V$  unique words in the entire corpus. In that case, the size of the TF-IDF matrix will be 2-dimensional  $D \times V$ . TF-IDF matches words with a statistic that makes those words diagnostic for identifying a certain document or set of documents, by rewarding a document/page/search return object for the frequent use of uncommon words while punishing documents for frequently using common words in general. Straightforward TF-IDF is considered a fairly simple document-ranking model, but many (perhaps even a majority of) ad-hoc search systems on the internet today use TF-IDF or a refinement of it.

An ideal semantic vector space has been trained on every possible felicitous input from a given language. Of course, while the number of felicitous inputs is basically infinite, the point of diminishing returns is only reached in corpora (such as those in English, Spanish, Russian,

or Chinese) of enormous size. TF-IDF is only as good as the documents you use to build it – there is no real “training” process, just very large amounts of first-order matrix arithmetic. More advanced “word embedding” models that work with higher dimensions have to be trained on very large amounts of text in the target language, to develop some kind of internal sense of a distributional “vocabulary” into which they can insert the new documents. The necessary amount of data is not available in Tibetan, though we have found that word embeddings with a high-resource language baseline (like English in our spaCy model) can and do still perform better than strict language-internal matrix generation methods like TF-IDF.

When a text is transformed into a dense embedding vector, it not only interprets the text as a sequence of numbers but also positions that sequence in conceptual space. It is thus envisaged as existing in multiple (mathematical) dimensions. In the case of the Cohere multilingual model (one of the types of embeddings offered by Cohere), it is a space with 768 dimensions. By projecting the vectors in this multidimensional space, the computer can subsequently perform a range of algebraic operations (angle comparison → cosine similarity, dot product, etc) to measure similarity, compare them analogically, and so on. In simple language, because the computer can recognise that two numbers or numerical arrays are close to each other in a mathematical sense, it can infer that once those strings have been translated back into the words or tokens they represent, those words or tokens will be similar in meaning. This is the basic principle behind semantic searching.

Both topic modelling and semantic searches require some capacity for dynamic updates – meaning the capacity for the machine to improve its knowledge over time as new resources are added to the corpus. This is what a possible data flow (pipeline inputs) for the creation of a useable vector space for topic modelling and/or searching might be:

*Raw text → Tokenised Text → Selected Features →  
Embeddings // Understanding Space*

Each of the steps following “Tokenised Text” broadly relies on increasingly complex matrix algebra, but the critical input unit for any of those linear-algebraic operations is a “feature”, a term generally used to refer to any word in a text apart from the stopwords. Put more simply, generally speaking (though not always), a feature is always a word, but not every word is a feature. Those selected features will then be vectorised, although the specific shape or dimensionality of the vector array varies by method. Stopword removal is not necessary for LLM-type searches, but it is for traditional Boolean-keyword searches<sup>3</sup> and various common topic modelling methods like LDA.

### 1.2 *Word Embeddings*

Word embeddings were the technological successor to document vectorisation and were industry-standard from the mid-2010s (Word2Vec, for example, was released in 2013) until the large-scale development of transformers from 2017 onwards. Canonical methods of word embeddings at that time included Word2Vec and later GloVe (Global Vectors for Word Representation, developed by the StanfordNLP group). Word2Vec and GloVe are optional embedding methods included in boilerplate spaCy, and both embedding models are critically reliant on large training corpora for ideal performance. Word2Vec represents terms (words, tokens,  $n$ -grams) as “dense” vectors in a single space generated by a reading of the entire corpus at once to generate the distributional semantics of features, and the size of the embedding space is determined by the specific embedding method or formula. Dense matrices fill all the cells of their matrices with meaning-bearing values, unlike sparse matrices (such as TF-IDF), where each row and/or column corresponds to a single feature or unit of analysis. Overwhelmingly, most cells in a sparse matrix will thus

---

<sup>3</sup> A Boolean keyword search uses logical operators (AND, OR, NOT) to combine search terms and narrow or expand results. For example, “cats AND dogs” yields results containing both terms, while “cats OR dogs” finds results containing either term. NOT excludes terms, like “pets NOT snakes” to find pet content without snake-related results.

have value 0, because a sparse matrix element at position  $ij$  will only be non-zero when the row index of  $i$  is the same as the column index of  $j$ .

Unlike TF-IDF, in a dense embedding matrix,<sup>4</sup> the size of the final space is not dependent on the number of documents or unique words, but rather leverages a series of dimensionality reductions as needed in order to shrink or grow the size of the space to the user's preset specification. The previous generation of Tibetan NLP research (see, for example, Tao *et al.* 2020) invoked Word2Vec as its preferred embedding framework. Unfortunately, because the Word2Vec Tibetan model is closed and only its product (output) is publicly released, then we as end-users would have to retrain a Word2Vec model, which would also mean having to reinvent the wheel for tokenisation, and so forth. If one has a good tokeniser that can be integrated into a very simple pipeline, generating TF-IDF matrices is much easier than training a model to develop a new embedding space for a previously unknown language.

At the time of writing, only corporate entities like Cohere have developed embedding spaces for Tibetan that are practically useful and publicly available. Cohere's model architecture is proprietary and not publicly available but certainly leverages some kind of transformer model (the same technology underlying LLMs like OpenAI's GPT) to create better context-sensitive embeddings for input text. It should also be noted that, from version to version, Cohere's multilingual engine varies in its ability to handle data across languages: Cohere Multilingual Model 2.0 had a much better "natural" understanding of Tibetan than the current available model, 3.0, so we continue to use Model 2.0 for our embeddings. Being a corporate entity, Cohere does not publicly release information about their data sources or model training process; it is difficult, perhaps impossible to speculate why their newer model might have worse Tibetan understanding than the older one. Our communications with the Cohere engineering team suggests that even they are unsure of the details that led to a drop in performance on some languages with the newest version of their

---

<sup>4</sup> Note that dense embedding matrices have more dimensions.

multilingual model. Our experience with the Cohere embedding space has been positive, and we recommend its use to others interested in a variety of embedding-driven NLP tasks for Tibetan, especially searching and document classification.

## 2 *Topic Modelling and Semantic Searching*

Our project is mainly interested in using digital methods to identify “divergent discourses” – changing topics or narrative foci over time in a set of texts. We therefore need to use topic modelling – automatic, corpus-scale identification of textual foci – with Tibetan texts. Topic modelling and its uses are discussed in more detail in Schwartz & Barnett 2025 in this issue, so here we give a simplified description of its basic principles in order to show how it differs from semantic searching. Topic modelling identifies topics or themes in a text based on algebraic forms of detection – i.e., it creates mathematical abstractions of the co-occurring vocabularies of a text to detect narrative threads, whether or not the explicit name of a “topic” is mentioned. These themes are extracted based on the semantic similarity of content-bearing texts. As a result, the identified topic or theme may be labelled with a word that never occurs in the text it is attributed to. One approach to topic modelling is that used by the iLCM. It infers topics using LDA, a conditional probabilistic method. LDA assumes, firstly, that documents are collections of “topics,” which it understands as inherently mathematical abstractions corresponding roughly to narrative threads, and secondly, that topics are composed of words, and topics can be extracted from corpus-level rather than document-level word distributions, the former of which of course are too large to be visible to an end-user. It uses statistical (Bayesian) methods to calculate how diagnostic or indicative certain word combinations are of a new topic and then plots the distribution of those content words to identify probable topics. The user can adjust or refine that probability by predefining a few likely topics, if so desired. LDA requires that you specify ahead of time how many topics to search for. Most importantly, it is at its core “keyword” based, in

that all the “words” on which it works are listed ahead of time and in effect searched for. LDA can identify multiple topics per document and is based on word frequency, working in a somewhat similar way to building concordances to determine whether Shakespeare wrote a play.

For some purposes, a different approach to topic modelling will be more relevant, depending on what one is trying to accomplish. As we have seen, transformer models are the current standard for high-performance NLP applications. Transformers achieve improved contextual understanding over previous-generation NLP tools mainly by encoding individual tokens based on the encodings of other tokens in the same sentence, so that polysemous and homophonous distinctions will naturally emerge given enough training data (a “bow” for shooting arrows, bending at the waist, wrapping a gift, and the front of a ship would each be assigned unique representations), while the model also stores positional information about each token to learn and reproduce good natural language syntax.

For a task such as topic modelling, transformers, when combined with other methods, have proven effective when used with high-resource languages. Among these transformer-hybrid methods, BERT (bidirectional encoder representations from transformers), a language model developed for generating embeddings of texts (Devlin *et al.* 2019) was applied to Tibetan texts by scholars in Tibet in 2022.<sup>56</sup> At that time, this approach was primarily used only for document classification. BERTopic, a topic modeller based on the BERT architecture, is perhaps the best known (Grootendorst 2022). Ronald

---

<sup>5</sup> Although it has since been outpaced by stronger models, BERT remains the standard against which large language model architectures are measured.

<sup>6</sup> Two BERT models for Tibetan were released in 2022: TiBERT (Sun *et al.*, Minzu University) and Tibetan BERT (Zhang *et al.*, Tibet University). The two models mainly differ in size and computational complexity: TiBERT is an adaptation of the baseline BERT model,<sup>6</sup> and was designed with text classification in mind. Tibetan BERT is freely available online through HuggingFace, while TiBERT is available through the creators' own distribution, and includes neither the training data nor a detailed description of it. These represented an early revolution in transformer technologies for Tibetan NLP but have largely been outmoded by the more potent forces in corporate AI services.



Schwartz has since succeeded in developing BERTopic as a tool for topic modelling with modern Tibetan texts (Schwartz & Barnett 2025).

Topic modelling allows a user to conduct “outside-in” analysis of a corpus, such as looking at what topics tend to occur over time within a corpus or identifying what is represented in a particular article. Textual analysts, however, also need an “inside-out” method, where they know what kinds of things they want or expect to see but do not know where to find them. For this we need a search system that can match “meanings”, or informal (rather than modelled) topics broadly considered, independent of previously assembled keywords. This is the task we call semantic searching. Semantic searches do not exactly do topic modelling – they find documents (or normally paragraphs) that are similar to a given query text, but they do not give those documents a set of semantic labels. Instead, they keep them in their “natural” state. In a user-driven search application, a semantic search pipeline vectorises the text query input by the user and then compares its similarity with that of the vectorised corpus. It then ranks the documents in the corpus according to their similarity to the query. This leads to overall better search results than simply searching for a specific keyword, since semantic search will detect words that are lexically dissimilar but have semantically similar values or contexts, as with “Tibetologists” and “people who study Tibetan texts”. In the same way, if the word “military” is in the query, semantic search will also return documents about “army” and “navy”. The results of a semantic search are thus based on similarity of meaning rather than word frequency or resemblance.

### 3 *Building a Semantic Search Tool for Tibetan*

The possibility of developing semantic search for Tibetan, and the use of Cohere to create such a system, was first proposed by Ronald Schwartz in 2023 (Schwartz & Barnett 2025). Our communications with various members of the Tibetan studies community indicate that he was the first to realise that such a system could be developed for Tibetan and that the tools needed for this task were by then already

available. As a pilot project, he first developed a semantic search system for use with newspaper texts in Simplified Chinese. These were based on the Cohere multilingual model and formed the architectural basis for the Tibetan semantic search tool. The Schwartz prototype digested the Chinese-language versions of his test corpus, which consisted of several thousand *Tibet Daily* articles stored in HTML files, covering years from 2008–2023. Those HTML files were then converted into CSV files using a bespoke parser that separates the main text into chunk “paragraphs” of roughly 256 characters. These were then associated with their metadata. His tool also demonstrated the practicability of providing toggleable automatic translation in English for each Chinese chunk or paragraph.

Working closely with Schwartz, we then adapted the basic architecture of his Chinese-language semantic search engine for use with texts in modern Tibetan. We again used Cohere, since it is unique among AI services in providing built-in support for Tibetan. We needed to resolve two issues in particular: an accessible form of storage for our vectors, and a supply of data in Tibetan as our test corpus. To store our vectors (for each document there is a vector of 768 numbers, totalling ~278,000 vectors for our test corpus), we used a cloud storage service called Pinecone, which specialises in storing vectors and provides an API to access the vector database for end-users to create their own search engines. To test our semantic search engine, Schwartz provided four years of newspaper articles in Tibetan that he had previously collected between 2020 and 2024. The articles are in HTML form and are remarkably consistent in format and easily parseable but first had to be preprocessed for use by the search tool. This we did by writing a new bespoke parser to extract the text in paragraph-sized chunks (which can be detected by line breaks). The parser also extracts metadata from the HTML file, including the title of the article, the reported author, the date of publication split into its long date string plus its individual components, the publication source and other features, such as the issue number and the source file name.

Using these articles as our test corpus and Cohere to create numerical representations or embeddings, we were then able to build a semantic search system around the embeddings, storing the index in

Pinecone. This index includes not only the vectors but also the raw text and the metadata. New texts can be added dynamically to the corpus by vectorising them with the same method that generated the initial embeddings. New queries can be treated in the same way, except that queries are not saved to the corpus. When we add a new article, the system needs to calculate how similar it is to all existing articles. While this takes more time as our database grows, the processing time increases in a manageable way (linearly rather than exponentially). This is because adding a new article only requires calculating its relationships to existing articles - it does not affect how existing articles relate to each other.

A full schematic of our search system structure can be found in Figure 1. Note that circles indicate objects, squares indicate operations, and green boxes are bespoke python programs that necessarily interact with their connected processes. Coloured circles are interactable from the user's perspective; white circles are hidden from the user's direct access unless specifically declared.

The "raw text with metadata" corresponds in our case to the CSV containing the information we parsed from the HTML files, but the exact format of the data is not of great importance as long as it is machine-readable. The text column of this raw data is then chunked into smaller pieces if and only if the length of the object in the text column is greater than 256 syllables – that is, in effect, if the sum of the text + (!) *shad* + punctuation marks is greater than 256, since Cohere's tokeniser tends to interpret every Tibetan syllable as (computationally speaking) a token. The chunked text is then embedded (converted into vectors) and stored in the database using the Pinecone script, which encircles the embedding process and includes a toggle for inspecting the vectors when they are produced, in addition to sending them to the Pinecone database. When the vectors appear in the Pinecone database, they can also be manually inspected (mainly for the correct dimensionality; the array of numbers itself is inscrutable to humans). This process generates the vector database. Each paragraph in the CSV will have an associated ID in the vector database that corresponds to

its vector. The paragraph ID is associated to the CSV table, which is what allows the comparison operation to be searchable.

To interpret the query, we follow an essentially identical process: the

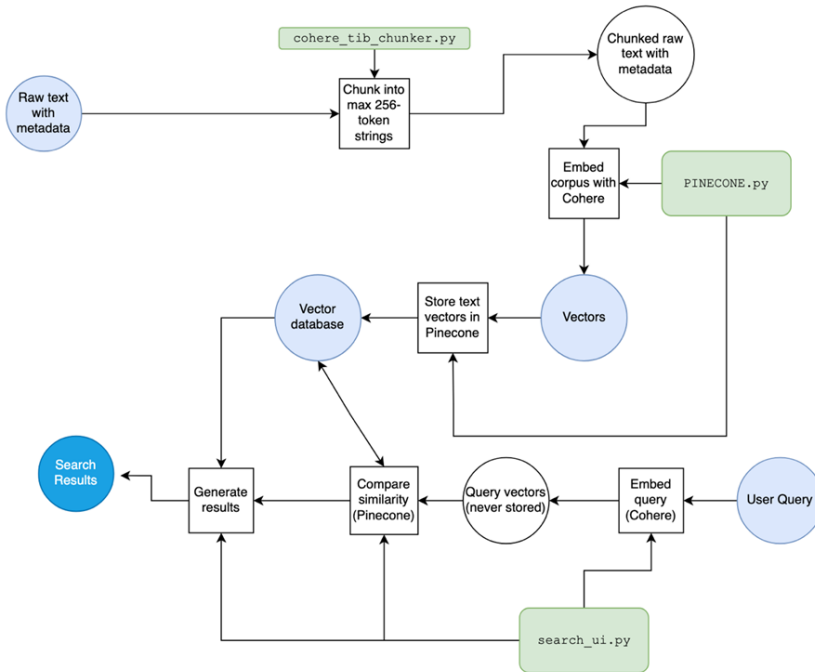


Figure 1 Search System Schematic

query is dynamically embedded in exactly the same understanding space as the other vectors (note that it is *essential* to use Cohere Multilingual Model 2.0 here, and not Multilingual 3.0). The embedded query is then measured against the embeddings of the paragraphs in the corpus using cosine similarity (not inner dot product), and the  $n$  most similar results are returned to the user. All parts of the query pipeline in our prototype are handled in an integrated system with the user interface. After the query results are returned, the query embeddings are forgotten by the system; the query embedding pipeline must be kept separate from the corpus in order not accidentally to contaminate the database with endless queries.

We also added several external functionalities to our search interface: a number of metadata fields (dates, source, title of article,

and so on) and a simple user interface that allows the user to rank results either by relevance (semantic match) or by metadata fields. We added a toggle allowing non-readers of Tibetan to access a translation of either the query or the results in English. For this, we integrated Bing Translate, which, at the time of writing, appears to be the most reliable provider of online translations from Tibetan to English. The public link to the search tool is <https://tibetcorpus.uni-leipzig.de/search/>.

A sample query and result from the semantic search engine is shown in Table 1. The query is taken from an online Tibetan-language news report about the Tingri earthquake in January 2025. The result that the search tool found in our test corpus to be most similar in terms of meaning was an article from a July 2020 issue of *Tibet Daily*. The query refers to the importance of early warnings regarding aftershocks and other risks following the earthquake, and the need to send relief forces to the site of the disaster rapidly. The result does not include the word earthquake or any similar term, but it describes a meeting four years earlier which had stressed the importance of almost exactly the same measures listed in the query. It is a paragraph (or <256 tokens) in length, so it includes additional information besides that included in the query, such as references to epidemic control. But it mirrors, in different words, all the main points in the query about the importance of early warnings and the rapid dispatch of relief forces in the event of what we can guess from context is an earthquake.

Table 1 Sample of semantic search and top result

<p>Query</p>	<p>ཡམ་འཕྲོ་དང་རི་ཉིལ་བ་སོགས་ལ་ལྷ་ཞིབ་ཚད་ལེན་དང་སྤྲོན་བརྗེ་གཏོང་བར་ཤུགས་སྤྲོན་རྒྱག་པ། གནས་གཤིས་འགྲུར་སྤྲོག་ལ་དྲ་སྤུར་ཚུན་པོ་བྱེད་པ། ས་གཤིས་གཞོན་འཚེའི་མངོན་མེད་རྒྱུན་དུ་ཡོད་པ་ཡོངས་ཁྲུབ་གྲུབ་ལ་བཤེར་བཅས་བྱས་ཏེ་ཞོར་ཐོན་གཞོན་འཚེ་སྤྲོན་འགོག་ནན་པོ་བྱེད་དགོས། ཟུང་དང་། གློག་ལས་སོགས་རྩལ་གཞིའི་སྤྲིག་བཀོད་ལ་གཏོར་སྤྲོན་པོ་གཤམ་རྩལ་ཤུགས་ཡོད་རྒྱུ་སྤུར་བཟོ་བྱེད་པ་དང་གོགས་སེལ་སྤོབས་ཤུགས་དང་གོགས་སེལ་དགོས་མཁོ་ཚན་རིག་དང་མཐུན་པའི་སྤྲོན་ས་བཀོད་གཏོང་བྱེད་པ་དང་ཡམ་འགོག་གོགས་སེལ་གྱི་རྒྱ་རྒྱས་འཁོར་དང་། དེའི་རྒྱུ་ལེན་ལ་ཤར་སྤྲོད་ཐུབ་པའི་འགན་ལེན་བྱེད་དགོས།</p>
<p>Bing Translation</p>	<p>Strengthen monitoring and early warning of aftershocks and landslides, pay close attention to climate change, survey geological hazards, and strictly prevent the occurrence of secondary disasters. Where roads and other infrastructure facilities are damaged, disaster relief forces and disaster relief</p>



decreases as the language of the query moves further from major global languages – our impression of the quality of results from a Vietnamese query was not favourable. But we found that, in terms of similarity, entering a query in Chinese or English produces accurate results from the corpus in Tibetan. The multilingual functionality of the system, both in query parsing and translation of results, means the tool can be used by a general audience, is not limited to readers of Tibetan, and does not require any programming knowledge. Our user interface for the search system, which is hosted on a private university server, is available for use by the public and includes a few useful design elements: sliders for restricting searches in years and months, toggleable translation of both queries and results, and a toggleable option for machine-readable and human-readable results. Figure 2 is a screenshot of the homepage with the same query as in Table 1:

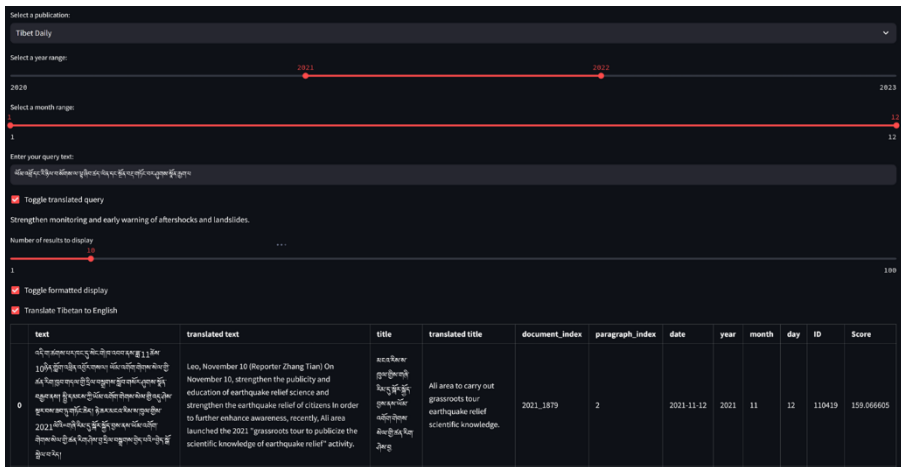


Figure 2 Semantic search homepage with query

We aim to add all digitised corpora from the Divergent Discourses project to the searchable system, which will provide the first ever wide-scope system for searching historical Tibetan newspapers – or any other set of Tibetan documents – using state-of-the-art NLP tools. Further, by integrating the spaCy tokenizer developed by Kyogoku *et al.* 2025 (in this issue), we can append context-sensitive keyword search functions to the semantic search function already developed. Integrating keyword and semantic search will dramatically simplify

the arduous process of searching for specific passages in digitised Tibetan corpora. Search systems like this are not computationally intensive to build and do not require extensive knowledge of software design, and to do so will only get easier as language models for Tibetan improve.

Finally, we ought to draw attention to a few reasonable improvements to the tool. The most obvious is to convert the interface to a proper web-based language like JavaScript and deploy it securely on a standard website. We also want to integrate a keyword search functionality, like TF-IDF, alongside the semantic search system in the same interface, to allow users to easily switch between short and long queries, and between identical and similar results. We additionally hope to implement the ability to subselect returned results and re-sort them, filtering or removing unwanted results from the initial search. The system costs roughly US\$2 per month to maintain, between server costs, API requests, and usage-based subscriptions to services like Bing Translate; higher traffic will naturally tend to increase those costs, though not into outrageous amounts. However, because the understanding spaces are properties of external corporate entities, and because those entities change, recalibrate, decommission, or close off their current services or models, it is not possible to be certain how long access and functionality will remain available. As NLP solutions for low-resourced languages like Tibetan gravitate to transformer-based technologies, the reliance of these solutions on external services and corporations could make their long-term sustainability increasingly uncertain.



## Bibliography

Blei, David M., Andrew Ng, and Michael I. Jordan

“Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3, 2003. pp. 993–1022. Available online at <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> (accessed January 26, 2025).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.

“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *North American Chapter of the Association for Computational Linguistics*, 2019. Available online at <https://arxiv.org/pdf/1810.04805> (accessed January 15, 2025).

Faggionato, Christian and Marieke Meelen

“Developing the Old Tibetan Treebank.” In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Varna: INCOMA, 2019, pp. 304-312. [doi:10.26615/978-954-452-056-4\\_035](https://doi.org/10.26615/978-954-452-056-4_035)

Grootendorst, Maarten

“BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure,” *arXiv preprint*, 2022. [doi:10.48550/arXiv.2203.05794](https://doi.org/10.48550/arXiv.2203.05794).

Hackett, Paul G.

“Automatic Segmentation and Part-Of-Speech Tagging For Tibetan: A First Step Towards Machine Translation.” In *Proceedings of the 9th Seminar of the International Association for Tibetan Studies*, 2000, pp. 1-18. Available online at <http://hdl.handle.net/10022/AC:P:10471> (accessed on December 18, 2024).

Jiang Di

“A New Perspective for Modern Tibetan Machine Processing and its Development: an Insight into the Method of

Computerized Automatic Understanding of Natural Languages in Terms of Chunk Identification." In Xu, B., MS. Sun, and GJ. Jin (eds.) *Some important problems in Chinese language information processing*. Beijing: Science Press of China, 2003, pp. 438–448.

Kyogoku, Yuki, Franz Xaver Erhard, James Engels, and Robert Barnett  
 "Leveraging Large Language Models in Low-resourced Language NLP: A spaCy Implementation for Modern Tibetan",  
*Revue d'Etudes Tibétaines* 74, 2025, pp. 187–222.

Meelen, Marieke, Élie Roux, and Nathan W. Hill  
 "Optimisation of the Largest Annotated Tibetan Corpus Combining Rule-Based, Memory-Based, and Deep-Learning Methods," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 20 (1), 2021, pp. 1-11. [doi:10.1145/3409488](https://doi.org/10.1145/3409488).

Schwartz, Ronald, and Robert Barnett  
 "Religious Policy in the TAR, 2014–24: Topic Modelling a Tibetan Language Corpus with BERTopic," *Revue d'Etudes Tibétaines* 74, 2025, pp. 221–260.

Spärck Jones, Karen  
 "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentation* 28 (1), 1972, pp. 11–21. [doi:10.1108/eb026526](https://doi.org/10.1108/eb026526).

Sun Yuan, Liu Sisi, Deng Junjie, and Zhao Xiaobing  
 "TiBERT: Tibetan Pre-trained Language Model," *arXiv preprint*, 2022. [doi:10.48550/arXiv.2205.07303](https://doi.org/10.48550/arXiv.2205.07303).

Tao Jiang, Li Jia, Ma Cao Wan, and Jia Hao Meng  
 "The Text Modeling Method of Tibetan Text Combining Word2vec and Improved TF-IDF." *Journal of Physics: Conference Series* 1601, 2020. [doi:10.1088/1742-6596/1601/4/042007](https://doi.org/10.1088/1742-6596/1601/4/042007).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, and Aidan N. Gomez

“Attention is All You Need,” *Advances in neural information processing systems* 30 (1), 2017, pp. 261–272. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762).

Zhang, Jiangyan, Deji Kazhuo, Luosang Gadeng, Nyima Trashi, and Nuo Qun


“Research and Application of Tibetan Pre-training Language Model Based on BERT.” In *Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics*, 2022, pp. 519–524. [doi:10.1145/3548608.3559255](https://doi.org/10.1145/3548608.3559255).



# *The Tibet Mirror*, “Friends” of Tibet, and the Internationalisation of the Tibet Question<sup>1</sup>

Natalia Mikhailova

(Université Laval, SOAS University of London)

 ssued from 1925 to 1963 in the Indian city of Kalimpong by Dorje Tharchin Babu (1890–1976), *The Tibet Mirror* has gained fame as one of the few early Tibetan periodicals and as the only newspaper which consistently promoted pro-Tibetan nationalist views in the Tibetan language in the 1950s.<sup>2</sup> Modern Tibetan nationalism can be said to have emerged first among the Tibetan-speaking elite and then, a little later, among the rest of the population of the Tibetan Plateau when they fully realised the determination of the new communist government in Beijing to integrate all Tibetan peoples and their territories into the big multinational “family” of China. Building on George Dreyfus’s claim that “to a large extent Tibetans did not have a full-fledged nationalism before 1950,”<sup>3</sup> this paper argues that *The Tibet Mirror* was the first Tibetan-language media organ to promote a modern form of Tibetan nationalism, serving as an important instrument for shaping a pro-Tibetan

---

<sup>1</sup> The author would like to express sincere gratitude and acknowledge the generous support of her research provided by the Fonds de recherche du Québec (FRQ) and Université Laval.

<sup>2</sup> For more details, see Moskaleva 2016, 2018, 2020.

<sup>3</sup> As an example of Tibetan proto-nationalism, Dreyfus (2005: 10–11) cites the unifying narratives found in the collection of texts known as the *Maṇi Kambum* (*ma ṇi bka’ ’bum*), which describe the special role of the Bodhisattva Avalokiteśvara as a protector of the Tibetan people and establishes his connection to the Tibetan king, Srong btsan sgam po (c. 605–650).

nationalist ideology and educating Tibetans on how the Tibet Question might best be presented to the world.

Print media is often termed “the fourth branch of power,” after the legislative, executive, and judicial branches, and is seen as able to shape public thought in major ways by implanting particular ideas, images, associations, and stereotypes in the minds of readers and thus directing the attitude of the readership towards a specific subject or even completely altering their world outlook (Danilova 2014: 11–12). The editor of *The Tibet Mirror* was clearly aware of the power vested in his hands (cf. Moskaleva 2020: 433) and, from the very early 1950s onwards, regularly published articles on Tibetan independence in his newspaper. He enthusiastically contended that Tibet had been independent for centuries, that Tibet would have still been independent if it had not been for the “vicious”<sup>4</sup> politics of the “treacherous” Chinese communists,<sup>5</sup> and that Tibet would become independent again in the future. In *The Tibet Mirror*, Tharchin presented antagonistic images of the Other (i.e., the Chinese communists) as the evil aggressor and the Self (i.e., the population of the Tibetan Plateau) as its helpless victim. In his writings of the 1950s and early 1960s, the editor set forth the elements of a national narrative which is now widespread within the Tibetan diaspora. He suggested concrete steps to be taken to pursue the goal of Tibetan independence, including direct appeals to rise against the Chinese communist forces and to resist them with arms, as well as described more intricate ways to secure Tibet’s independence. Among the modes of action that Tharchin proposed was the acquisition of support for the Tibetan cause from the international community and the cultivation of foreign “friends” of Tibet, whether individuals or entire countries, who, in his view or in that of “learned” experts, favoured the idea of an independent Tibet and provided help to the people of Tibet either in word or deed, or both.

In the following pages, I will use selected abstracts from *The Tibet Mirror* from the 1950s and early 1960s to illustrate Tharchin’s attempts

---

<sup>4</sup> *sdug po* (*The Tibet Mirror* (TIM), Vol. XXI, No. 8, Nov. 1953, p. 2).

<sup>5</sup> *ngan g.yo can dmar lugs 'dzin pa* (TIM, Vol. XXI, No. 5, Aug. 1953, p. 6).

to establish the category of Tibet's "allies" and his strategies to articulate the Tibet Question most effectively.

## 1 "Friends" of Tibet

### 1.1 Great Britain

One of the earliest examples of Tharchin's focus on foreign friendship from the period under review can be found in *The Tibet Mirror* issued on September 1, 1950. In that issue, Tharchin published two brief statements expressing thanks to two British "friends" of Tibet. In the first of these two articles, entitled "Support for Tibetan Independence,"<sup>6</sup> the editor gave credit to the Graeco-British scholar Marco Pallis (1895–1989):

The esteemed Mr. Marco Pallis, [also] known by his spiritual name Thub bstan Bstan 'dzin, who has acquired a perfect knowledge of the beautiful ancient traditions of the Tibetan Religious State and the teaching of the Buddha, recently spread the [following] information in English newspapers: "If all countries, having negotiated their support for the independence of the Tibetan Religious State, find a way to leave [Tibet] as an ornament of the world, it will surely be very beneficial for all countries of the world."

[I would like to] express my gratitude for this [public statement] and thank him.<sup>7</sup>

In the second statement, Tharchin expressed his gratitude to the British diplomat Arthur Hopkinson (1894–1953):

The esteemed Hopkinson, the renowned former Political Officer in Sikkim, also spread the word via radio and newspapers that Tibet was independent. I am very much obliged [to him] as well.<sup>8</sup>

Tharchin described both "friends" of Tibet in a rather flattering manner and in this way reinforced the authority of their opinions on

<sup>6</sup> *bod rang btsan la brgyab skyor* (TIM, Vol. XVIII, No. 10, Sept. 1950, p. 9).

<sup>7</sup> TIM, Vol. XVIII, No. 10, Sept. 1950, p. 9.

<sup>8</sup> TIM, Vol. XVIII, No. 10, Sept. 1950, p. 9.

Tibetan independence. Arthur Hopkinson had limited himself to a brief remark on Tibetan independence, but Marco Pallis, according to Tharchin, with his “perfect knowledge of the beautiful ancient traditions of the Tibetan Religious State and the teaching of the Buddha”<sup>9</sup> was well acquainted with Tibetan history and culture and had appealed to other countries in the world not simply to support the idea of Tibetan independence, but to treat Tibet as “an ornament of the world.”<sup>10</sup> Tharchin, however, did not elaborate as to how or why this gesture would, as Pallis claimed, “surely be very beneficial for all countries of the world.”

Both Hopkinson and Pallis were connected with British intelligence. Hopkinson, the last British Political Officer in Sikkim, Bhutan, and Tibet (in office 1945–1947), had worked as the British Trade Agent in the third largest town in Tibet, Gyantse (Rgyal rtse), from 1927 to 1928 (McKay 1995: 263–264). In the words of Alex McKay, “trade was of little concern” to British Trade Agents; their main task was to cultivate friendly relations with Tibetans and to collect relevant information about the political and socio-economic situation on the Tibetan Plateau (McKay 2001: 94). The so-called “Trade Agents” belonged to the diplomatic corps of the Government of British India and were under the direct control of the Political Officer for Sikkim, Bhutan, and Tibet who was based in the Sikkimese capital, Gangtok (McKay 2001: 94). The Political Officer for Sikkim, Bhutan, and Tibet was directly responsible for British relations with these territories and oversaw the British Mission in Lhasa during his visits to the Tibetan capital (McKay 1995: 1, 270). Hopkinson and the editor of *The Tibet Mirror* collaborated in the context of Tharchin’s work for British intelligence (Fader 2009: 181–184, Sawerthal 2018: 132, n. 188).

As for Pallis, he was known for his fondness of mountaineering and for his interest in Tibetan Buddhism. According to the elder brother of the 14<sup>th</sup> Dalai Lama, Gyalo Thondup (Rgyal lo don ’grub, b. 1927), Pallis was so “deeply immersed in the study of Tibet” that he used to

---

<sup>9</sup> *bod chos ldan rgyal khab kyī gna’ srol bzang po dang sangs rgyas kyī bstan par yid kyis rab tu bsten.*

<sup>10</sup> *’dzam gling mdzes brgyan du ’jogs thabs gnang.*

dress in a Tibetan *phyu pa* (Rgyal lo don grub 2015: 168). Pallis was also associated with British intelligence: Gyalo Thondup reports that in the early 1960s, it was Pallis who arranged his meeting with the head of MI6, the British Secret Intelligence Service, in London (Rgyal lo don grub 2015: 168). The involvement of Pallis in intelligence activities in connection to Tharchin and financial support for *The Tibet Mirror* has been corroborated by Anna Sawerthal (2018: 109–111, 121, 124, 267).

In McKay's opinion, the political status of Tibet did not play a significant role for the British, but it was important for them that Tibet stood as a stable northern frontier of British India and helped in securing an effective barrier against Britain's rival in Central Asia—Russia (McKay 2001: 97). Therefore, until India gained its independence in 1947, British colonial officials supported the traditional Tibetan government of the Dalai Lama and encouraged the Ganden Phodrang (Dga' ldan pho brang) with its aim of strengthening and unifying Tibet so that it could claim "its place among the world's nation-states" (McKay 2001: 97). In the 1950s and 1960s, a few former British colonial officials who had previously served in Tibet attempted to provide more publicity for the Tibet Question (McKay 1995: 254). However, regardless of their efforts, both before 1947 and after, the British government never officially recognised Tibet as an independent state and, on the contrary, sought to avoid damaging Sino–British relations and so did not express overly explicit support for Tibet. Tsering Shakya contends that after India gained independence, Great Britain completely lost interest in Tibet and left the Tibet Question to the discretion of independent India (Shakya 2000: 19).

Nevertheless, despite Britain's withdrawal from Tibetan politics after 1947, Tharchin continued to present a positive image of the British in *The Tibet Mirror* in the 1950s. For example, in November 1950, the editor reported that representatives of the Tibetan delegation were "pleased" by the British government's promise to provide them with "the best possible assistance"<sup>11</sup> when they would be traveling through

---

<sup>11</sup> *in gzhung nas grogs ram gang drag gnang rgyu* (TIM, Vol. XVIII, No. 12, Nov. 1950, p. 6).



Hong Kong to Beijing for Sino–Tibetan negotiations.<sup>12</sup> Tharchin also frequently repeated in his publications that the British colonial government had participated in the negotiations in Simla and that Great Britain had recognised in the Simla Convention that Tibet was not under the Chinese rule but was “merely in the shadow of China”:<sup>13</sup>

When the agreement was signed in Simla in 1914, Great Britain recognised that Tibet was merely in the shadow of China. However, the Chinese government did not object to that. According to the agreement, the Chinese government was not allowed to control Tibet, to take over [Tibet], to expand [its territory in Tibet], or to interfere in anything [in Tibet]. Now [November 1950], the Tibetan government has ordered its representatives to negotiate [with the PRC] precisely on the basis of that earlier agreement [reached] in Simla.<sup>14</sup>

After a large number of Tibetans and Tibetan-speaking people followed the 14<sup>th</sup> Dalai Lama into exile in India, Great Britain was still presented in Tharchin’s publications as Tibet’s “friend.” For instance, in February–March 1961, the editor published a report entitled “The British Government’s Help to Tibetan Refugees,”<sup>15</sup> in which he publicised the financial support of benevolent British “friends” of Tibet:

Recently, on March 17 [1961], an esteemed ambassador of the British government in India, Sir Paul Gore-Booth,<sup>16</sup> handed over to his colleague K.L Mehta, the Secretary General of the Indian Foreign Ministry, the help of the British government for Tibetan refugees who had arrived in India in the form of 50,000 English pounds, which is equal to 666,000 Indian rupees.<sup>17</sup>

---

<sup>12</sup> *in ji nas grogs ram gnang rgyu’i bka’ mol la ‘thus mi rnams thugs mnyes po byung ‘dug* (TIM, Vol. XVIII, No. 12, Nov. 1950, p. 6).

<sup>13</sup> *in ji nas bod rgya nag gi grib ‘og tsam du ngos len gnang ‘dug.*

<sup>14</sup> TIM, Vol. XVIII, No. 12, Nov. 1950, p. 6.

<sup>15</sup> *bod kyi skyabs bcol ba rnams kyi ched du dbyin gzhung gi mthun rkyen* (TIM, Vol. XXVII, No. 6, Feb.–Mar. 1961).

<sup>16</sup> In the publication, Tharchin slightly distorts his name as “Sir Paul Gore Broth.”

<sup>17</sup> TIM, Vol. XXVII, No. 6, Feb.–Mar. 1961, p. 6.

1.2 *India*

In the metanarrative of Tibetan independence constructed in *The Tibet Mirror* in the 1950s and 1960s, India appeared as another old “friend” of Tibet. Tharchin portrayed India as the “country of Bodhisattvas”<sup>18</sup> with a “religious government,”<sup>19</sup> which logically entailed its essential affinity to the “Tibetan Religious State.”<sup>20</sup> In the 1950s, the editor regularly mentioned the subject of close cultural and religious ties between Tibet and India, but not between Tibet and China. For example, in September 1954, Tharchin emphasised the impressive centuries-old connections between India and Tibet in the spheres of religion, science, and trade and elaborated on the richness of the bonds of friendship and mutual assistance between the two countries.<sup>21</sup> After the uprising in Lhasa in 1959, Tharchin persisted with even greater enthusiasm to promote the narrative of the longstanding Tibetan–Indian friendship. This is evident in an “abridged summary of a rough translation of news”<sup>22</sup> that Tharchin published in April 1959, based on a report that had appeared in the Calcutta-based newspaper *The Statesman*, on March 31 that year.

That report (see Appendix 1) summarises a long speech by Nehru, or at least Tharchin’s interpretation of it. One notices first the emphasis placed on the strength and closeness of historical ties between India and Tibet; these are alluded to three times in the article. Nehru is cited as arguing that such ties exist between India and Tibet at present (“India and Tibetans have strong kinship, cultural, and other ties”<sup>23</sup>), that India designs its Tibetan policy while being guided by these ties (“based on its close longstanding cultural and religious ties with Tibet,

---

<sup>18</sup> *rgya gar 'phags pa'i yul* (TIM, Vol. XXV, No. 12, May 1959, p. 5).

<sup>19</sup> *rgya gar chos ldan gzhung* (TIM, Vol. XXIII, No. 12, Nov. 1956, p. 3).

<sup>20</sup> *bod chos ldan rgyal khab*, the term that was predominantly used by Tharchin to refer to Tibet.

<sup>21</sup> TIM, Vol. XXII, No. 5, Sept. 1954, p. 4.

<sup>22</sup> *gnas tshul rags bsgyur mdor bsdu*s (TIM, Vol. XXV, No. 11, Apr. 1959).

<sup>23</sup> *rgya gar dang bod mi rnams gyis bar lal\_nye tshan gyi 'brel ba dang shes rig sogs kyi 'brel ba brtan por yod*.

[India] feels profound sympathy for Tibetans”<sup>24</sup>), and that India wishes to preserve these ties with Tibet (“we want to maintain close relations and pure mutual friendship with Tibetans”<sup>25</sup>).

In Tharchin’s account, Nehru criticises the PRC for using physical force against Tibetan monks (India sees “the damage which has been brought by communist China to the numerous monastic communities as a result of its unprincipled violence and which has become the reason for the [total] decay of good virtues [there]”<sup>26</sup>) and “urges” China to “expand” the “freedom and independence”<sup>27</sup> of Tibet. As if echoing the opinion of *The Tibet Mirror* editor himself, Nehru, according to Tharchin, declares that the Tibetan government has never recognised its subordination to China.<sup>28</sup> Thus, with respect to the Sino–Tibetan confrontation, India in *The Tibet Mirror* fully supports Tibet, even though formally—“in terms of politics”—its hands are tied<sup>29</sup> and at the same time, India would even like to maintain its friendly relations with the PRC.

Tharchin’s account also includes a smoothing of the brutal “legacy” of colonial India and Great Britain in Tibet. During Younghusband’s military expedition to Tibet, British soldiers used modern artillery weapons and machine guns to kill hundreds of primitively armed Tibetans, plundered monasteries and forts, and forced the Tibetan government to sign a treaty on conditions that were entirely favourable for the British (Powers & Templeman 2012: 271, 740–741; McKay 2012: 14, 19). However, in Tharchin’s rendering, Nehru mitigates these facts by presenting Younghusband’s expedition to Tibet as “not a daunting intervention with a takeover sanctioned by

---

<sup>24</sup> *bod dang sngon nas shes rig dang chos sogs kyi 'brel gnas dam por yod par brten/\_bod mi rnams la sha tsha'i sems tshor chen por yod.*

<sup>25</sup> *nga tshos bod mi rnams dang mthun 'brel gtsang mar nye 'brel byed 'dod yod.*

<sup>26</sup> *rgya dmar nas tshul min dbang gis dgon sde khag mang gtor skyon btang ba de ni gang min bsod nams nyams pa'i rgyu red.*

<sup>27</sup> *nga tsho khong tshor rang dbang rang btsan yar 'phel yong rgyur yid 'dod chen po zhu gi yod.*

<sup>28</sup> *rgya nag gzhung re re nas bod 'di rgya khongs yin par brjod 'dug kyang /\_bod gzhung rim pa nas ngo len byas mi 'dug.*

<sup>29</sup> *rgya gar nas khrims bstun ltar bod skor srid phyogs nas gang yang byed mi thub.*

the government's order."<sup>30</sup> Nehru is also described as claiming in his speech that India, even if it had seized Tibetan territories earlier, had always renounced its rights to them. Yet, according to Tsering Shakya, in October 1947, two months after India gained its independence, the Tibetan government demanded that India return the Tibetan territories that had been "gradually annexed by the British," and Nehru dismissed the demand as unrealistic (Shakya 2000: 280). Later, the unresolved territorial disputes on the border of India and Tibet, further fuelled by the activities of Tibetan guerilla fighters in the late 1950s and early 1960s and the diplomatic strain after the flight of the 14<sup>th</sup> Dalai Lama to India, contributed to piling up tension in the region and ended up in the form of the Sino-Indian War of October-November 1962.

Later statements in *The Tibet Mirror* attributed to the Indian Prime Minister on the Tibet Question were more reserved and diplomatic in their tone. For example, in July 1959, Tharchin included another excerpt from Nehru's speech, but unlike three months earlier, this time the Indian Prime Minister said nothing regarding the political status of Tibet and confined himself to a summary of the current conditions offered for Tibetan immigrants in India:

As it was reported in the news from New Delhi on July 7, [1959], during a monthly meeting with journalists in Delhi, the esteemed Prime Minister Nehru made a speech.

Esteemed Prime Minister Nehru [said] regarding the venerable Dalai Lama: "It is difficult to distinguish what must be said, [but] we clearly see how he is doing. We are equally aware of [his] feelings [because of] the difficulties. The venerable Dalai Lama has reached the age of 25.<sup>31</sup> Not only he [himself], but his fellow countrymen too have lived through a terrifying experience.

---

<sup>30</sup> *'di ni bkas bskos rgyud 'dzin pa'i dbang gnon the byus byas pa ma red.*

<sup>31</sup> The 14<sup>th</sup> Dalai Lama Tenzin Gyatso (Bstan 'dzin rgya mtsho) was born on July 6, 1935, which means that in 1959 he was 24 years old according to the Western system of age-counting. Nehru, in Tharchin's account, is apparently using the East Asian system of age-counting, in which one year is added to the child's age at birth. Tsering Shakya reports that 1959, when the 14<sup>th</sup> Dalai Lama turned 25, was considered to be astrologically unfavourable for the Tibetan leader, which was why some Tibetans believed that 1959 would bring difficulties not only for the 14<sup>th</sup>

As you [can] understand, since he is the main religious leader [of Tibet], we paid him our respects with a warm welcome. As long as the venerable Dalai Lama is in India, we will treat him with great love and respect. [We] have not only welcomed about 12,000 more people [from Tibet] but, having gradually divided the Tibetans who asked for asylum into groups, we will send them to various mountain areas. The refugees will not be sent to the plains. [Some of them] will be sent to the ancient monasteries of Sikkim, and some will be settled in other areas unrelated to Sikkim. The children of [Tibetan] refugees have been and are being sent [now] to schools and similar institutions. We are trying to overcome these difficult circumstances gradually.”<sup>32</sup>

In Tharchin’s narrative from July 1959, India gives “a warm welcome”<sup>33</sup> to the 14<sup>th</sup> Dalai Lama and the Tibetans who fled after him, takes into consideration the high-altitude factor of Tibetan native areas while arranging the refugee settlements and tries to resolve the existing difficulties. Nehru expresses his sympathy and empathy for the people of Tibet and the Dalai Lama, yet he makes it clear that Tenzin Gyatso (Bstan ’dzin rgya mtsho) has the status of a “religious leader,”<sup>34</sup> not a political one. The Indian Prime Minister describes the experience of the Dalai Lama as “terrifying,”<sup>35</sup> but diplomatically refers to Tibetan refugees as the Dalai Lama’s “fellow countrymen,”<sup>36</sup> but not as his “subjects,” or his “people.”

Before publishing this report, Tharchin included a one-sentence statement in his paper: “Any kind of Tibetan government on the Indian territory will not be recognised.”<sup>37</sup> No matter how much the editor of *The Tibet Mirror* would have hoped for the opposite, but this time Nehru confined his discourse strictly within the framework of

---

Dalai Lama himself, but for all Tibetans (Shakya 2000: 191). Perhaps, Nehru mentions the 14<sup>th</sup> Dalai Lama’s age in his speech in connection with this prejudice popular among Tibetans back at the time.

<sup>32</sup> TIM, Vol. XXVI, No. 1, June 1959, pp. 7–8.

<sup>33</sup> *dga’ bsu zhus*.

<sup>34</sup> *chos kyi dbu gtso*.

<sup>35</sup> *’jigs rung gi nyams myong myangs pa red*.

<sup>36</sup> *khong gi lung pa’i mi rnams*.

<sup>37</sup> *rgya gar sa gnas su bod gzhung dang ’dra zhig yin rung ngos ’dzin byed kyi ma red* (TIM, Vol. XXVI, No. 1, June 1959, p. 7).

Indian humanitarian aid, which Tharchin could not present as support for Tibetan independence.

In the July 1959 issue of *The Tibet Mirror*, Tharchin allocated more space to promoting positive images of India and Sikkim as “friends” of Tibet who readily came to the aid of Tibetan migrants (see Appendix 2). According to Tharchin, India comes to the rescue and does “everything possible”<sup>38</sup> for Tibetans who fled the PRC: it grants asylum to all refugees, tries to settle them in cooler areas, improves their living conditions, collects donations for them, and takes good care of Tibetan lamas and monks. Sikkim,<sup>39</sup> being portrayed as another important “friend” of Tibet, also provides tangible and no less sympathetic assistance to Tibetan immigrants. At the same time, Tharchin underlines in his report that “what matters most” is that, despite the difficulties, “the venerable Dalai Lama and the Tibetan government will certainly make every possible effort”<sup>40</sup> for the benefit of Tibetans. The editor of *The Tibet Mirror* thus supports the construction of a narrative in which the Tibetan Ganden Phodrang government, having migrated along with the 14<sup>th</sup> Dalai Lama to India, did not lose its legitimacy and continued working for the benefit of its subjects, with the only difference being that now it was to perform its duties from the territory of India.

Tharchin’s publications about India’s concern for the Tibet Question were not limited to reports about Indian humanitarian aid for Tibetan refugees. In the late 1950s and early 1960s, the editor also made room in his newspaper for Indian public figures who condemned Chinese communist oppression and human rights violations in Tibet. For instance, in November–December 1959, under the resonant title, “A Drama of Suffering and Terror Is Being Played in Tibet,”<sup>41</sup> Tharchin published a translation of a public statement on the Tibet Question made by the Indian politician Kanaiyalal Maneklal Munshi (1887–1971; see Appendix 3).

---

<sup>38</sup> *thabs shes gang drag gnang mus su mchis*.

<sup>39</sup> The Indologists Yurlov and Yurlova (2010: 313) define Sikkim of that period as “a quasi-sovereign state which was a protectorate of India.”

<sup>40</sup> *gtso che gong sa mchog dang bod gzhung nas thabs shes gang drag gnang nges*.

<sup>41</sup> *bod du mya ngan dang 'jigs skrag can gyi zlos gar zhig khrab bzhin pa red*.

Munshi’s speech was a highly critical assessment of the Chinese communist policy in Tibet, expressing views similar to Tharchin’s anti-communist position. Munshi, an activist of the Indian independence movement, politician, and writer, is presented in *The Tibet Mirror* as speaking on behalf of all India. His statement was published nine months after the Dalai Lama’s flight to India, but it is much harsher in its rhetoric than Nehru’s speech the previous April. In Munshi’s rendering, there is no longer any talk about Chinese suzerainty, and instead, the emphasis is placed on the fact that Tibet had certainly been independent (“earlier Tibet used to be an independent religious state just like us”<sup>42</sup>). Additional ambivalence is added when speaking about the years since 1950 when Tibet was already officially part of the PRC, Munshi implies that Tibet still nevertheless held its “independent” political status: “The introduction of any malevolent and highly improper measures [aimed] at the restriction of human rights and of the religious sphere of an independent [Tibetan] state are nothing but indecent evil acts which violate ethical norms.”<sup>43</sup>

Much as Tharchin had done in his own writing, the Indian politician portrays the PRC as an aggressor and a tyrant. In three consecutive sentences, Munshi uses grammatical constructions to suggest that the Chinese aggression will be of indefinite duration and aimed not only at Tibet but at the entire continent of Asia: Chinese communists “have conducted and are conducting a campaign of violent expansionism,”<sup>44</sup> “have said and are saying that they will liberate not only Tibet, [but] all people of Asia,”<sup>45</sup> and “have unbearably oppressed and are [currently] oppressing Tibet.”<sup>46</sup>

---

<sup>42</sup> *sngon du bod 'di ni nga tsho dang 'dra bas chos ldan rang btsan rgyal khab gcig yin.*

<sup>43</sup> *rang btsan rgyal khab de'i mi'i dbang tshad dang /\_chos phyogs su shin tu mi mtshams pa'i lag len ngan pa ji dang ji bstar ba rnam ni mi chos khrims 'gal gyi las ngan 'tshabs chen sha stag bgyis.*

<sup>44</sup> *rgya nag han zer ba'i mi rigs de dag ni rgya 'gyed btsan bshed ring lugs kyi las 'gul byas dang byed mus red.*

<sup>45</sup> *de tshos bod tсам ma yin par e she ya'i mi rigs yongs la bcing bkrol btang gi yin zhes brjod dang brjod mus red.*

<sup>46</sup> *rgya dmar nas bod la mnar gcod mi bzod pa btang dang gtong bzhin.*

Munshi emphasises twice the “terrifying situation” of the “unbearable suffering”<sup>47</sup> “and the “terrible situation in Tibet,”<sup>48</sup> while accusing the PRC of “the extermination of the fine ancient traditional religion and culture of the entire Tibetan nation, as well as the social norms, rights, and basic welfare of [all] human beings;”<sup>49</sup> and “destroying the entire Tibetan nation.”<sup>50</sup> Tharchin’s account of the speech concludes with an assurance of India’s sympathy for Tibet and Munshi’s concern for the Tibet Question: “It is important to help [lay] the foundation of peace in Tibet.”<sup>51</sup>

It was not only Munshi who contributed to the construction of the image of the PRC as a dangerous neighbour “at the doorway”<sup>52</sup> of India. Tharchin actively promoted this idea himself and thus tried to insert some alienation into Sino-Indian relations. For instance, in the February–March 1961 issue, he published an ominous warning, with a poetic ending, entitled “Fears That the Happiness of Minority Nationalities of the Himalayas Is Going to Be Devoured by Red Demons from the East”<sup>53</sup> (see Appendix 4).

In the article, Tharchin introduces the conspiracy theory of the “Five Fingers,” which are assumed to be of critical importance to India and which, in his opinion, are going to be seized<sup>54</sup> by the PRC to undermine the “strength”<sup>55</sup> of India. The conception of the foreign policy theory called “The Five Fingers of Tibet (西藏的五指 *xizang de*

<sup>47</sup> *bod yul du deng sang mya ngan mi bzod pas mnar ba'i gnas lugs 'jigs su rung ba'i skor thos.*

<sup>48</sup> *rgya dmar nas [...] bod la gnas lugs ngan pa rgya cher spel ba red.*

<sup>49</sup> *bod rigs yongs rdzogs kyi sngar srol bzang po'i chos dang shes rig rnam dang 'gro ba mi'i tshad dbang bde rtsa rnam rtsa med bzo 'gyur brtsams pa red.*

<sup>50</sup> *bod rigs yongs rdzogs rtsa med bzos pa red.*

<sup>51</sup> *bod la zhi bde 'thob pa'i gzhi rtsar phan grogs nges par gnang dgos gal che zhes gsungs 'dug.*

<sup>52</sup> *rang re's sgo 'gram la.*

<sup>53</sup> *hi ma la ya'i ri bsul du chags pa'i mi rigs grangs nyung rnam kyi bde skyid la shar phyogs srin dmar gyis za sems la dwogs zon.*

<sup>54</sup> *rgya gar gyi byang phyogs su lag sor lta bu'i sa gnas 'gag rtsa che ba nub stod la dwags dang smad shar a sam khul dang / bar du khyim mtshes rgyal khab bal yul dang 'bras 'brug khag la [...] za sems dbang gnon byed phyir dmar po'i srid byus spel thabs dang / rgya gar nas tha dad du 'phral thabs byed kyi yod pa red.*

<sup>55</sup> *rgya gar gyi stobs shugs la gnod thabs byed kyi yod pa red.*



*wu zhi*)” is attributed to Mao Zedong. Still, no official information has been found to confirm that. According to the theory, the leader of the PRC held Tibet as the “palm” of a “hand,” while the territories of Nepal, Bhutan, Sikkim, Ladakh, and present-day Arunachal Pradesh were the five “fingers” of it.<sup>56</sup> After the incorporation of Tibet into the PRC in 1951, Chinese communists allegedly sought to create a buffer zone on the Sino–Indian border, which the so-called “five fingers” were to form (Singh 2013: 1; Smith 2013: 27). By promoting the “Five Fingers of Tibet” theory, Tharchin, first and foremost, tries to antagonise India against the PRC, while also provocatively urging Sikkim, Nepal, and Bhutan to “stay very cautious” and “not get fooled”<sup>57</sup> by the promises of the Chinese.

The editor uses the label “demons”<sup>58</sup> for Chinese communists, literally excluding them from “the human realm,”<sup>59</sup> where representatives of the PRC cause “the real experience of suffering.”<sup>60</sup> To characterise the policy of “communist demons,” Tharchin uses graphic similes (the 17-Point Agreement is “like a razor smeared with honey and a hat made of wet leather;”<sup>61</sup> “harmed” India is “like [...] a palm without fingers which is left without any strength”<sup>62</sup>), as well as metaphors (owing to the fault of “demons,” Tibetans suffer from “committing suicide with the razor [smeared with honey]”<sup>63</sup> and having their heads “squeezed ... by the hat made of wet leather”<sup>64</sup> that

---

<sup>56</sup> Singh 2013: 1. The professor of the US Naval War College Paul J. Smith argues that the author of the “five fingers” metaphor was the British journalist Desmond Doig, who published an article entitled “India to Protect Border States” in *The Washington Post* on Aug. 26, 1959 (Smith 2013: 27, 34). It is interesting that the publications on the “Five Fingers” theory by both Teshu Singh and Paul J. Smith came out in 2013.

<sup>57</sup> *de la dwogs zon chen po gnang zhing / dbu ma 'khor na ha cang yag po red.*

<sup>58</sup> *srin mo; srin dmar.*

<sup>59</sup> *mi yul.*

<sup>60</sup> *sdug bsngal [...] dngos myang ji byung la gzigs.*

<sup>61</sup> *sbrang rtsi byugs pa'i spu gri dang klo rlon zhwa mo dang 'dra ba'i gros mthun don mtshan 17pa.*

<sup>62</sup> *sor mo med pa'i lag mthil la nus shugs gang yang bral ba ltar gyaa gar [...] la gnod.*

<sup>63</sup> *spu gris rang srog bcad.*

<sup>64</sup> *ko rlon zha mos rang mgo btsir.*

starts to dry up), and epithets (“feigned love,”<sup>65</sup> “pleasant demonic lies”<sup>66</sup>). At the same time, the editor repeats four times that Tibet was and still is an “independent state”: “earlier, there was an independent Tibetan Religious State between India and China,”<sup>67</sup> “the independent Tibetan state stood like a border guard,”<sup>68</sup> “the independent Tibetan government submitted a petition,”<sup>69</sup> and “the independent Tibetan state has no choice but [...]”<sup>70</sup>

Concluding the article with four seven-syllable lines, Tharchin resorts to hyperbole to declare the destruction of “the entire Tibetan nation”<sup>71</sup> and uses syntactic and lexical parallelism to liken the Five Principles of Peaceful Coexistence<sup>72</sup> to the 17-Point Agreement, in which Tibet officially recognised itself as part of the PRC in 1951. Thus, the editor tries to convince India that China deceived it in 1954 and seeks to create further tension in Sino–Indian relations.

### 1.3 *The Kuomintang*

Who else, in Tharchin’s opinion, was among the “friends” of Tibet? In the 1950s and 1960s, representatives of the Kuomintang—unlike Chinese communists—were described in *The Tibet Mirror* not as “enemies” of Tibet, but rather as allies of Tibetans. Tharchin referred to the Kuomintang as the “true” or “genuine” Chinese government<sup>73</sup> and, most importantly, claimed that the Kuomintang government

---

<sup>65</sup> *bcos ma'i brtse ba.*

<sup>66</sup> *rgya dmar srin mo'i g.yo gtam snyan po.*

<sup>67</sup> *sngon du rgya gar dang rgya nag gnyis kyi bar la bod chos ldan rang btsan rgyal khab yod.*

<sup>68</sup> *bod rang dbang rgyal khab kyis [...] sa srung ba lta bur gnas yod pa red.*

<sup>69</sup> *bod rang btsan gzhung gis [...] snyan gseng zhus.*

<sup>70</sup> *rang btsan bod rgyal khab de yang da cha rgya dmar btsan 'og tu mi tshud ka med byung ba red.*

<sup>71</sup> *bod rigs thams cad phung.*

<sup>72</sup> The principles known in Hindi as the Panchsheel (पंचशील) were adopted by India and the PRC in 1954 in the Preamble to the Sino–Indian Agreement on Trade and Intercourse between the Tibet Region of China and India (Tikhvinskiy 2017: 570).

<sup>73</sup> E.g., see TIM, Vol. XXVI, No. 1, June 1959, suppl. 2 or TIM, Vol. XXI, No. 5, Aug. 1953, p. 7.

would grant Tibet independence if it regained its control over mainland China.<sup>74</sup> In the Chinese news column in *The Tibet Mirror*, the editor advocated for the return of the Kuomintang to mainland China and predicted that Chiang Kai-shek would be successful in his confrontation with the CCP.<sup>75</sup>

Tharchin repeatedly placed news about the Kuomintang alongside news from Tibet. For example, on the front page of *The Tibet Mirror* in the January 1954 issue, amid news about the Kuomintang, the editor published a letter from Tibet in which the author criticised the policy of the PRC and the “liberation” of Tibet by the PLA.<sup>76</sup> Directly above this letter, Tharchin placed a piece of news stating that the Kuomintang government was about to attack the Chinese Communist Party. Immediately after the letter, the editor shared the news that the government of the US was currently assisting the Kuomintang with arms and that Kuomintang troops would very soon be seen entering mainland China.<sup>77</sup>

In the 1950s and 1960s, Tharchin portrayed the Kuomintang as a valuable ally of Tibetan nationalists not only because any “enemy” of the PRC automatically was considered a “friend” of defenders of Tibetan independence, but also because, in Tharchin’s interpretation, the alliance with the Kuomintang was virtually equivalent to the alliance with the US. The editor recurrently emphasised that the Kuomintang enjoyed the broad support of the United States and that, therefore, the coalition of interests between these two governments left the PRC with no chance to win. For instance, in the September 1954 issue, under the title “Exaggeration regarding the Liberation of Formosa, or Taiwan,”<sup>78</sup> Tharchin argued that any threats to Taiwan’s security made by the Chinese communist government would meet an equally strong response from the U.S. government—if the PRC pursues its aim to “certainly liberate” Taiwan,<sup>79</sup> the United States “will

---

<sup>74</sup> TIM, Vol. XXVI, No. 1, June 1959, suppl. 2.

<sup>75</sup> TIM, Vol. XXVI, No. 1, June 1959, suppl. 2.

<sup>76</sup> TIM, Vol. XXI, No. 10, Jan. 1954, p. 1.

<sup>77</sup> TIM, Vol. XXI, No. 10, Jan. 1954, p. 1.

<sup>78</sup> *phor mo sa'am da'i wan bcings bkrol gtong rgyu'i 'ur gtam.*

<sup>79</sup> *ring ming phor gling bcings bkrol byed nges gtan.*

certainly defend" it.<sup>80</sup> He uses the word "certainly" twice, but at the same time indicates that his account is an approximation ("replied something like this"<sup>81</sup>):

Beijing radio recently announced: "Soon, we will certainly liberate the island of Formosa. No foreign state is allowed to interfere at that time. If anyone does so [i.e., interferes], he will get into trouble."

The American President replied something like this: "The 7th Marine Regiment [of the US] will certainly defend the island of Formosa."

When I look at such statements, I clearly see that if the [Chinese] Communist Party invades the island of Formosa, the United States will provide support to Chiang Kai-shek. Given the circumstances, would not the conflagration of the Third World War get ignited at Formosa, as a result of which the soldiers and communist fire[power] would destroy themselves and others?<sup>82</sup>

On the same page of *The Tibet Mirror*, immediately following this news, under the title "Liberation of China,"<sup>83</sup> Tharchin reported on the decisive plans of the Kuomintang leader himself. Here, he depicts the Chinese communists as oppressing the people of China and as the "enemies of the Buddhist teaching,"<sup>84</sup> while Chiang Kai-shek, who is identified as "the genuine ruler of China,"<sup>85</sup> is said to promise "certainly" not just "liberation from the oppression of the followers of communist Russia,"<sup>86</sup> but also independence:<sup>87</sup>

The genuine ruler of China, Chiang Kai-shek, said: "All people of China that are [still] left in the isolated areas will soon be certainly made happy as a result of liberation from the oppression of the followers of communist Russia."

<sup>80</sup> *phor gling la mtsho dmag ang bdun pas bsrung skyobs byed nges gtan yin.*

<sup>81</sup> *de'i lan lta bu.*

<sup>82</sup> TIM, Vol. XXII, No. 5, Sept. 1954, p. 11.

<sup>83</sup> *rgya nag bcings dkrol.*

<sup>84</sup> *bstan dgra gung bran* (e.g., TIM, Vol. XXVII, No. 6, Feb.–Mar. 1961, p. 5).

<sup>85</sup> *rgya gzhung ngo ma'i spyi khyab cang kai shag.*

<sup>86</sup> *btsan dbang nas rgya nag yongs kyi mi dmangs mya ngam thang lus rnam la bcings bkrol thog bde la 'god rgyu nges gtan yin.*

<sup>87</sup> *slad phyin bod rang dbang rang btsan du 'jog rgyu'i rtsa 'dzin sogs kyang gtan la phab bsgrubs.*

Not only did he [Chiang Kai-shek] say that, [but] he said that Tibetan Buddhism, which is being forcefully harmed by [the spread of] atheism in the religious region, will also be liberated. It is said that even a program, according to which Tibet will be left independent in the future, was settled.

In any case, even though both sides claim that they will liberate each other, at present, one can only wonder who will set whom free. If [the Kuomintang] does not liberate [Tibet], then the Third World War will liberate [it].<sup>88</sup>

Throughout the 1950s and 1960s, Tharchin repeatedly published articles intended to imbue his readers with a sense of trust in the benevolent intentions of the Kuomintang towards Tibet. For instance, *The Tibet Mirror* issue of July–August 1958 said that the Kuomintang government had expressed “pleasure”<sup>89</sup> at news of the pro-independence rebellion in Kham, promised “to provide relevant large-scale assistance,”<sup>90</sup> and condemned Chinese migration to Tibet as intended to “destroy the Tibetan nation”:<sup>91</sup>

According to the rough translation of a report from a Chinese newspaper [published] in Calcutta, the Mongolian and Tibetan Affairs Commission in Formosa stated:

“[We] are very pleased that at present, the people in the Tibetan region of Kham, having gathered their own army, are organising a major uprising against communist China in order to protect their religious system and independence. The situation with their needs, goals, etc. will be analysed in detail, and in the future, the Mongolian and Tibetan Affairs Commission plans to provide relevant large-scale assistance. By relocating a large number of Chinese people to Tibet, communist China has been thus employing numerous strategies to destroy the Tibetan nation. Tibetans certainly need to be cautious of this.”<sup>92</sup>

<sup>88</sup> TIM, Vol. XXII, No. 5, Sept. 1954, p. 11.

<sup>89</sup> *deng dus khams bod khul du rang gi chos lugs dang rang dbang bsrung ched [...] rgya dmar la ngo rgol chen po byed kyi yod 'dug pa de la ha cang dga' spro chen po byung.*

<sup>90</sup> *gang la gang 'os kyi rogs ram rgya chen po byed rtsis yin.*

<sup>91</sup> *rgya dmar nas rgya yi mi dmangs mang po bod du btang nas bod kyi mi rigs med pa byed pa'i thabs byus mang po byed kyi yod pa red.*

<sup>92</sup> TIM, Vol. XXV, No. 3–4, July–Aug. 1958, p. 12.

After the 14<sup>th</sup> Dalai Lama's flight to India, Tharchin also covered the Kuomintang's assistance to Tibetan immigrants in his newspaper. For instance, in *The Tibet Mirror* issue dated November–December 1959, he published a “rough translation”<sup>93</sup> of an excerpt from a speech<sup>94</sup> by Li Yongxin (李永新 1901–1972), the official in charge of the Kuomintang's Tibet policy. According to the article, Li had expressed concern about the fate of Tibetan refugees (“the government of Taiwan needs to find a way to accommodate all of them”<sup>95</sup> and “the government of Taiwan will find a way to accommodate”<sup>96</sup> Tibetan refugees), referred to Tibetans twice as “brothers,”<sup>97</sup> claimed to be helping even those Tibetans who are not refugees but live permanently in Taiwan (“the government provides them with help, and their living conditions are very good”<sup>98</sup>), and offered to provide for other Tibetans who might decide to immigrate to Taiwan:

“After the uprising in Lhasa in March 1959, there are more than 25,000 Tibetan monks and laymen who went into flight. The government of Taiwan needs to find a way to accommodate all of them. Apart from that, more than 5,000 of them [Tibetan refugees] sent the [following] petition to the government of Taiwan: ‘Since we will certainly expel Chinese communists, [we] ask for help, support, and assistance.’ Therefore, in order to save Tibetan brothers, the government of Taiwan will find a way to accommodate [them],” —said [Li Yongxin].

Among Taiwan citizens, there are more than 700 people in the government who voluntarily help their Tibetan brothers. Moreover, people who donate money gave more than 800 Taiwanese dollars<sup>99</sup> to the government of Taiwan to help Tibet.

---

<sup>93</sup> *rags bsgyur*.

<sup>94</sup> “Head of the Kuomintang Mongolian and Tibetan Affairs Commission Li Yongxin Stated [the Following] at the Meeting,” (*go min tang gi bod sog las khungs kyi dpon po rlis yun shing nas tshogs 'dur gsungs par*).

<sup>95</sup> *de tshang ma da'i wan gzhung nas tshur blang thabs byed dgos*.

<sup>96</sup> *da'i wan gzhung nas [...] tshur blang thabs byed kyi yin*.

<sup>97</sup> *bod rigs spun zla rnams la rogs skyabs slad* and *bod rigs spun zla rnams la dwangs blang rogs ram byed*.

<sup>98</sup> *de tshor gzung nas rogs ram gnang zhing 'tsho gos sogs ha cang yag po 'dug*.

<sup>99</sup> Tharchin used the basic term *sgor* which can refer to any currency, but in context is assumed here to refer to Taiwanese dollars.

Nowadays, there are altogether 466 Tibetans and Mongols among the residents of Taiwan. The government provides them with help, and their living conditions are very good.<sup>100</sup>

Amid the continuous reporting on Chinese communists being the “enemies of the Buddhist teaching,” Tharchin also featured news on how the Kuomintang government made offerings to Tibetan monks during the Monlam Chenmo celebrations in India in 1961:

I heard the news that recently, during the Lhasa Monlam Chenmo Festival celebration at the Tibetan monastery in Bodhgaya, the Kuomintang government of China respectfully presented each monk with an offering of money, tea, and soup.<sup>101</sup>

Thus, in Tharchin’s anti-communist discourse, the Kuomintang government occupied an important position of a “friend” of Tibet and the Tibetan-speaking diaspora: the Kuomintang promised to grant independence to Tibet if it won over the CCP, it supported anti-communist uprisings in Tibet, its followers demonstrated a sympathetic attitude towards Tibetans residing in Taiwan, and—unlike representatives of the CCP—they expressed profound respect for Tibetan Buddhism.

Representatives of the Kuomintang did, indeed, offer financial aid and arms to the Tibetan insurgents (Shakya 2000: 170–172; Rgyal lo don grub 2015: 146), and in March 1959, pressured by the US Chiang Kai-shek declared the future prospect of Tibet’s self-determination if the Kuomintang succeeded in returning to mainland China (Shakya 2000: 231). However, the possibility of Tibetan self-determination under a future Kuomintang regime did not mean that the Kuomintang government had renounced the idea of Chinese sovereignty over Tibet. As the Japanese scholar Kensaku Okawa has noted, the official position of the Tibetan government-in-exile now regards any cooperation with the Kuomintang as compromising the claim of Tibet as independent from China, and Tibetans who cooperated with the Kuomintang government in the 1960s and 1970s are considered as “betrayers” (Okawa 2007: 607, 599).

---

<sup>100</sup> TIM, Vol. XXVI, No. 6–7, Nov.–Dec. 1959, p. 9.

<sup>101</sup> TIM, Vol. XXVII, No. 6, Feb.–Mar. 1961, p.6.

Tharchin's view of the Kuomintang as a "friend" of Tibet was, nevertheless, shared by some individual Tibetan politicians in exile. In particular, the 14<sup>th</sup> Dalai Lama's elder brother Gyalo Thondup—who as a youth had studied in China with financial support from Chiang Kai-shek—wrote in his memoir that Chiang Kai-shek treated him "as a son," "never said that Tibet had ever been part of China," and "was also willing for Tibet to remain independent" (Rgyal lo don grub 2015: 73–75). After the PRC had made clear its goal to "liberate" Tibet, Gyalo Thondup, who was married to a daughter of the leading general in the Kuomintang army (Knaus 2003: 64; Rgyal lo don grub 2015: 80, 121), went to Taiwan in May 1950 on a passport issued by the Kuomintang government and met with Chiang Kai-shek, who offered him the position of the Head of the Mongolian and Tibetan Affairs Commission (Shakya 2000: 40; Okawa 2007: 600; Rgyal lo don grub 2015: 118). After spending 16 months in Taiwan as a paid guest of Chiang Kai-shek, Gyalo Thondup left for the US with his Kuomintang-issued passport and a cheque for 50,000 US dollars from the Kuomintang leader (Rgyal lo don grub 2015: 121, 123). Only after his visit to the US did Gyalo Thondup change his mind on the favourability of cooperation with the Kuomintang and instead prioritise direct contact with the US. In his memoir, he explained his change of heart by noting the discrepancy between the Kuomintang's goal of fighting against Chinese communists to regain power in mainland China and "the struggle for a free and independent Tibet" by Tibetan fighters. In any case, in those years, he wrote, it seemed to him that "the US was so great and powerful that it could make almost anything happen" (Rgyal lo don grub 2015: 146).

It is also worth mentioning that in his memoirs, Gyalo Thondup notes that a group of Tibetan immigrants in Kalimpong who were working on the Tibet Question in India in the 1950s and 1960s (i.e., Gcen mkhan rtsis gsum<sup>102</sup>) "persuaded" Tharchin Babu to increase the

---

<sup>102</sup> The Gcen mkhan rtsis gsum was an anti-Chinese group of Tibetan *émigrés* formed in the 1950s in Kalimpong by Gyalo Thondup, the former Tibetan minister Shakapba, and the Tibetan monk official of the 4<sup>th</sup> rank Khenchung Lobsang Gyaltzen (Mkhan chung Blo bzang Rgyal mtshan). From 1954 onwards, the group unofficially cooperated with the Indian government and the Indian Intelligence



production of *The Tibet Mirror* to once a week (Rgyal lo don grub 2015: 149). Given that, the cooperation between Tharchin and Gyalo Thondup was probably not limited to a one-time contact about increasing the frequency of the newspaper’s production. Reading through the pages of *The Tibet Mirror*, one cannot help noticing that the views on the Tibet Question, certain wordings, and specific individuals mentioned by Gyalo Thondup in his book largely coincide with those mentioned by Tharchin in his newspaper in the 1950s and 1960s.

#### 1.4 *The United States*

References to the US in *The Tibet Mirror* are found not only in the context of the United States government’s assistance aimed at transforming Taiwan into an “unsinkable aircraft carrier”<sup>103</sup> that posed a threat to the CCP. In his newspaper, Tharchin also featured panegyrics acclaiming the splendour of the US in general. In May 1958,<sup>104</sup> he placed two articles facing each other on the same page. The first was headlined: “America Is the Good Mother of Asian Countries.”<sup>105</sup> To its right, the second headline read: “Communist China Has Become a Common Enemy of All Asian Countries.”<sup>106</sup> Whereas Tharchin labelled the PRC in *The Tibet Mirror* as a “common enemy”<sup>107</sup> or a “global enemy,”<sup>108</sup> he depicted the US as a “good mother” who was providing financial aid (in this article, it was US\$399 million) to various Asian countries (namely South Korea, Japan, Taiwan, the Republic of Vietnam, the Philippines, Thailand, and Laos).

---

Bureau as a pro-Indian Tibetan organisation and even received funding from India to gather intelligence on the situation in Tibet and on the Sino–Indian border (Goldstein 2014: 155–157, 161, 163, 168–169). For more details, see Goldstein 2014: 141–206.

<sup>103</sup> The metaphor is borrowed from Tikhvinskiy 2017: 569.

<sup>104</sup> TIM, Vol. XXV, No. 1, May 1958, p. 6.

<sup>105</sup> *a mi ri ka ni e she ya’i rgyal khab rnam kyi a ma bzang po zhig yin.*

<sup>106</sup> *rgya nag gung bran ni e she ya’i rgyal khab tshang ma’i spyi dgrar longs ’dug.*

<sup>107</sup> *spyi dgra.*

<sup>108</sup> *’dzam gling spyi’i dgra bo* (TIM, Vol. XVIII, No. 8, July 1950, p. 3).

It was thus protecting Asia, as he put it, from the “great danger” that Asian countries were certain to be exposed to if they were to “fall into the hands of communists.”<sup>109</sup>

Suggesting that the US shared something else—no less important—in common with the “Tibetan Religious State,” Tharchin introduced the concept of the US as the “American Religious State.”<sup>110</sup> Offering as visual evidence five photographs of the smiling Mongolian community in the US and of a new Buddhist temple that American citizens had built there for the Mongols,<sup>111</sup> Tharchin highlighted the supportive attitude of the US towards Buddhism and the potential for the establishment of friendly relations between the US and Tibet.

The positive image of the US as a “friend” of Tibet was also built on the basis of the American position on the Tibet Question and on their relations with the Kuomintang government. Tharchin tried to show in *The Tibet Mirror* that the US had great influence over the government of Chiang Kai-shek and that Americans supported Tibet’s independence, arguing that therefore, if the Kuomintang returned to mainland China, given their American support, Tibet was sure to become independent again. For example, in the April 1959 issue of *The Mirror*, Tharchin published “news from Washington”<sup>112</sup> that in his view clearly indicated US confirmation of the Kuomintang’s support for Tibetan independence:

The Press Secretary of the U.S. government said: “The esteemed leader of the Kuomintang government of China, Chiang Kai-shek, and his government repeated that since [they] treated with respect the religious and secular autonomous government of Tibet, they would be granting independence [to Tibet]. This fact made us happy.”<sup>113</sup>

Tharchin also gave space in his paper to news of America’s humanitarian aid to Tibetan refugees. In two news reports published immediately after the proclamation issued by “the Tibetan Volunteer

<sup>109</sup> *gal srid [...] gung bran gyi lag tu shor na e she ya rgyal khab tshang ma gnas thabs med pa'i nyen kha chen po zhig yod.*

<sup>110</sup> *chos ldan rgyal khab a me ri ka* (TIM, Vol. XXV, No. 7, Dec. 1958, p. 1).

<sup>111</sup> TIM, Vol. XXV, No. 7, Dec. 1958, pp. 1–2.

<sup>112</sup> *wa shing Ton gyi gsar gsal ltar.*

<sup>113</sup> TIM, Vol. XXV, No. 11, Apr. 1959, p. 3.

Army for Defence of Religion” in March 1959, the title of which was untypically provided in both Tibetan and English,<sup>114</sup> Tharchin first claimed that the US is ready to assist in the education of Tibetan youth and to arrange a permanent place of residence for Tibetan immigrants. In his second report, he detailed the humanitarian aid delivered to Tibetans by the US in the form of medicinal drugs and vitamins, worth in total US\$410,000.<sup>115</sup> Immediately after the report on the American aid, Tharchin published news that Taiwan had also provided assistance to Tibetan migrants and that the Tibetan government-in-exile had thanked the Kuomintang government for its help.<sup>116</sup> There was, however, no mention in *The Tibet Mirror* in the 1950s and 1960s of the CIA’s assistance to the Tibetan guerilla forces.

## 2 Internationalisation of the Tibet Question

While constructing the metanarrative of Tibetan independence in *The Tibet Mirror*, Tharchin not only shaped images of specific countries as “friends” of Tibet but also elaborated on the need for support of the Tibet Question from the international community. Tharchin discussed in this context the role of modern intergovernmental structures such as the United Nations (UN). As early as October 1, 1949, when the founding of the PRC was formally proclaimed, the editor tried to inform his Tibetan readership about the bases for obtaining membership of the UN. In an article entitled “Nepalese Government Attempts to Join the United Nations,”<sup>117</sup> he provided a “rough translation” of the submission of the Nepalese government to join the UN, which listed Nepal’s arguments for proving its independence as a state. These included the fact that it had sent diplomats abroad, that it had signed treaties with other nations (including Tibet), that it had its army, and that there were no foreigners in its government or army

<sup>114</sup> *bstan srid dmag sgar gyis dril bsgrags*; Eng. *Proclamation Made by the Tibetan Volunteer Army for Defence [sic] of Religion on 19-3-1959*.

<sup>115</sup> TIM, Vol. XXVI, No. 2–3, July–Aug. 1959, p. 15 [i.e. 17].

<sup>116</sup> TIM, Vol. XXVI, No. 2–3, July–Aug. 1959, p. 15 [i.e. 17].

<sup>117</sup> *gor Sha gzhung 'dzam gling mthun tshogs su tshud thabs*.

(see Appendix 5). Later, in September 1950, the editor mentioned this article from the previous October and gave his readers a sample of his reasoning for Tibet's admittance to the UN.<sup>118</sup> One can say that ever since the early 1950s, Tharchin had been consistently portraying the UN as another potential "friend" of Tibet, the attention that Tibetans needed to secure for their independence to be officially recognised by the international community.

The editor collected and presented in *The Tibet Mirror* any evidence of the international community's support or sympathy for Tibet in its fight against the Chinese communists. For instance, in June–July 1960, Tharchin reported that such an authoritative organisation recognised the oppression of Tibet by Chinese communists as the International Commission of Jurists:

A group of international jurists wrote down in a book reliable information about the Tibet Question and submitted [it] to the United Nations General Assembly. [They reported that] communist China had continuously oppressed and was [still] oppressing Tibet. In order to fully exterminate the Buddhist teaching, [Chinese communists] destroyed the protective stupas and other [sacred sites]. By means of "struggle sessions," [Chinese communists] killed many prominent lamas and *tulkus*, etc.<sup>119</sup>

The International Commission of Jurists (ICJ) was established in 1952 to counterbalance the International Association of Democratic Lawyers, which was under the control of the Soviet Union during the Cold War era and was thus perceived as a pro-communist organisation. However, the International Commission of Jurists was not an entirely independent and objective structure either, for the organisation was secretly sponsored by the CIA until 1967 (Tolley 1994: 31; Claude 1994: 576–577). Tsering Shakya describes the ICJ report on the situation in Tibet as "unashamedly pro-Tibetan" (Shakya 2000: 223). But, even though the organisation's report on Tibet reflected the strong anti-communist spirit of the 1950s and 1960s, the

---

<sup>118</sup> TIM, Vol. XVIII, No. 10, Sept. 1950, p. 5. For a full translation and analysis of the publication, see Moskaleva 2020: 420–423.

<sup>119</sup> TIM, Vol. XXVI, No. 12, June–July 1960, p. 10.

ICJ’s findings played a significant role in increasing publicity for the Tibet Question and in further internationalising it (Shakya 2000: 223).

Tharchin brought up the topic of internationalisation of the Tibet Question in many other articles. For example, in the August–September 1960 issue of *The Tibet Mirror*, under the title “Assembly of States of the World,”<sup>120</sup> the editor once again reminded his readers of the UN and the ICJ findings about Chinese communist policies in Tibet:

On the 20<sup>th</sup> of this month [September 1960], in the big American city called “New York,” there is going to be held the [15<sup>th</sup>] Session of the UN General Assembly. The Tibet Question is to be discussed during this meeting.

A group of international jurists carefully examined the Tibet Question. They presented a book which described in detail the destruction of the true Dharma by communist China and the complete annihilation of Tibetans.

Recently, the Indian government via its operational group<sup>121</sup> also submitted a petition to the [UN General] Assembly to request support for the Tibet Question.

When last year [1959], the Tibet Question was discussed during [the 14<sup>th</sup> Session] of the [UN General] Assembly, representatives of the Nepalese government not only did not support their neighbour Tibet, with which [the Nepalese] share a thousand years of mutual love for religion, but, furthermore, even initiated a harmful discussion<sup>122</sup> [on the Tibet Question]. Therefore, we Tibetans became greatly disheartened. If the Nepalese government has understood well what the policy of communist China is like, will they not provide support to the [discussion of the] Tibet Question this time?<sup>123</sup>

---

<sup>120</sup> *‘dzam gling spyi’i rgyal khab tshogs ‘du.*

<sup>121</sup> *ape ro e she yan tshogs pa.* Unfortunately, it was not possible to establish a precise translation of this term used by Tharchin.

<sup>122</sup> According to the results of the vote on the UN General Assembly resolution on the Tibet Question that was conducted on Oct. 21, 1959, Nepal abstained from the voting (“Question of Tibet: Resolution / Adopted by the General Assembly 1959”, *United Nations Digital Library*. Available online at <https://digitallibrary.un.org/record/664377?ln=en>, accessed July 24, 2024).

<sup>123</sup> TIM, Vol. XXVII, No. 1, Aug.–Sept. 1960, p. 12.

The motion to include the Tibet Question in the agenda of the 15<sup>th</sup> Session of the UN General Assembly did not, however, gain enough support from the UN member states—49 countries voted in favour, 13 against, and 35 abstained (Shakya 2000: 234). As a result, in 1960, despite the ICJ's report on the "annihilation" of Tibetans, neither the independence of Tibet nor the violation of Tibetan human rights by the PRC became the subject of resolutions passed by the UN that year.

While instructing the Tibetan readership on the importance of engagement with the UN in the struggle to regain Tibet's independence from China, Tharchin provided his readers with particular ideas for gaining more international support. These included strategies for successfully presenting the Tibet Question without even arguing about the political status of Tibet. On the last page of five successive issues of *The Tibet Mirror* published between June and October 1950,<sup>124</sup> Tharchin placed an advertisement in English for gramophone records and books teaching Tibetan language. The striking feature of these advertisements was not, however, the purchasable products, but Tharchin's rationale for the importance at that time of studying Tibetan. Continuing the British colonial tradition of the exoticisation of Tibet, Tharchin refers to Tibet as a "land of mystery,"<sup>125</sup> outlines its importance for world politics and emphasises its cultural heritage and its role as the "museum" of the world, protecting the cultural heritage of all Asia. Tharchin's narrative thus implies that the endorsement of Tibetan independence is essential for establishing and preserving peace in the whole world:

---

<sup>124</sup> See TIM, Vol. XVIII, No. 7, June 1950, p. 4; TIM, Vol. XVIII, No. 8, July 1950, p. 8; TIM, Vol. XVIII, No. 9, Aug. 1950, p. 6; TIM, Vol. XVIII, No. 10, Sept. 1950, p. 12; TIM, Vol. XVIII, No. 11, 1950, p. 6.

<sup>125</sup> For more information on the image of Tibet constructed by the British colonial officials, see McKay 1995: 189–198.

THE TIBETAN LANGUAGE CAN NOW BE LEARNT AT HOME  
WITHOUT A TEACHER<sup>126</sup>

Tibet is now rapidly gaining international fame and importance in world Politics. In fact the great powers of the world are taking more interest in Tibetan affairs. Tibet is not only a land of mystery but also the “MUSEUM” of the world, because it can claim to have preserved the ancient culture, religion and arts of Asia from ruin. The modern world can learn many things good from Tibet. Tibet has preserved vast ancient Sanskrit literature with faithful translation into Tibetan which are missing from India. It is, therefore, high time that India and all the powers of the world should take more interest in studying its religion, culture and various ancient manners and customs apart from its politics. One can learn many things from Tibet for bringing peace to the world. It is, therefore, very important that one should know the language well. This can now be done very easily with the help of a set of Tibetan language gramophone records and text books. The records are prepared by the Govt. of India and the text books by Sir B.J. Gould, C.M.G., C.I.E. the former Political Officer in Sikkim & Mr. H.E. Richardson, O.B.E., I.C.S. the present Officer Incharge of Indian Mission Lhasa, Tibet.

The text books and set of records are now available from THE TIBET MIRROR PRESS, KALIMPONG P.O. (W.Bengal.) India.<sup>127</sup>

As Alex McKay has argued, at the international level, the image of Tibet as a mysterious land contributed to the promotion of the idea that Tibet was a separate state (McKay 1995: 193). It also contributed to the image of Tibet as a cradle of Buddhism and a centre of spirituality (McKay 1995: 193–194). After 1959, this counterposing of the uniqueness of Tibetan culture and the exceptional spirituality of Tibetans with the atheism of Chinese communists was used by the Tibetan exile administration and its leaders as a strategy for winning international support for the Tibetan cause and the Tibetan diaspora (Brox 2006: 93). These efforts led to the culture of Tibet becoming so widely known across the world that it has been acknowledged as a

---

<sup>126</sup> Tharchin’s original orthography, syntax, and style are presented in the cited advertisement intact.

<sup>127</sup> TIM, Vol. XVIII, No. 9, Aug. 1950, p. 6.

part of global cultural heritage and as deserving of patronage by the international community (Brox 2006: 93). This in turn has helped the exile administration claim its status as the only legitimate representative of authentic Tibetan culture and of the genuine cultural and religious traditions of Tibetans (Brox 2006: 89).

Another important strategy used by Tharchin to internationalise the Tibet Question involved historical parallels. Danilova considered historical parallels as ways of manipulating the perception of discourse recipients by emphasising the external similarity of two compared objects or events, where the author “substitutes one phenomenon for another” and introduces a set of associations, emotions, and connotations that are relevant to the comparison but not necessarily to the reality of the compared objects (Danilova 2014: 84). This what Tharchin often did when he linked the Tibetan situation to precedents where other countries had either recently achieved international recognition as independent states or had received support of some kind from the UN during their struggles for independence.

One such article discussed the case of the former Outer Mongolia, which, like Tibet, had been part of the Qing Empire but had managed to gain its independence owing to the support of the Soviet Union.<sup>128</sup> Similarly, Tharchin repeatedly published vivid accounts of India’s fight for independence on the front page of his paper.<sup>129</sup> He also drew an analogy between the situation in eastern Tibet under Chinese communist rule and the Soviet–Yugoslav confrontation in 1948–1953. His source of inspiration was Josip Broz Tito (1892–1980), the then leader of Yugoslavia, who had become famous for initiating an independent political and economic development program for his country in the late 1940s. This had led to disagreement with Stalin and as a result Yugoslavia had been expelled from the Cominform; Soviet–Yugoslav relations remained severed until after Stalin’s death in March 1953 (Dvornichenko *et al.* 2008: 398–399). Shortly after, in July 1953, Tharchin published an appeal to the people of Kham to rise up

---

<sup>128</sup> TIM, Vol. XIX, No. 6, Sept. 1951, p. 6.

<sup>129</sup> E.g., TIM, Vol. XVIII, No. 10, Sept. 1950, p. 1.



“like Tito,” referring to them as the “Tito[ists]”<sup>130</sup> of Kham (see Appendix 6). In the article, Tharchin called three times for an uprising against “the Other”<sup>131</sup>—the communists—in eastern Tibet in the name of the “independence” of the entire Tibetan “motherland”<sup>132</sup> (“Rise up in the near future for independence like Tito!”).<sup>133</sup> Tharchin thus put Tibet, which had never been officially recognised as independent (or as including Kham), on a par with Yugoslavia, the independence of which had not been disputed since the end of the Second World War.

Later, Tharchin compared the fighting between the Khampas and Chinese communists in eastern Tibet to the suppression of the Hungarian Uprising by Soviet troops in 1956:

According to international law, a powerful state is not allowed to drop bombs from the sky and to fire from tanks on the ground into any protesters [who stand up] in the name of independence of their state, which is inferior in terms of strength and weapons. It is an illegal and evil act of the Chinese communists to have dropped bombs on the barely armed Tibetans who were practicing religion and standing up for the[ir] independence. It is reported that now, because of the fear of a recurrence of uprisings, [the Chinese communists] have additionally sent some tanks to Lhasa.

Recently, Russian communists carried out the invasion of Hungary in a similar manner by [bringing in] tanks and [dropping] bombs on people who stood up for their independence. However, it is reported that the people [of Hungary] continued resisting and fighting with even greater courage.

Recently, Great Britain and France invaded a republican country, and [in response] all states declared that it was unacceptable. However, there is probably not even one who would say “unacceptable” for the sake of our Tibet.

If there is no peace in Tibet, then there will probably be no peace in the whole world.<sup>134</sup>

---

<sup>130</sup> *kham bod khul du bzhugs pa'i ti to rnam.*

<sup>131</sup> *gzhan.*

<sup>132</sup> *rang ljongs rgyal khab rang btsan yun gnas ched.*

<sup>133</sup> *mi ring ti to bzhin du sger langgs shog.*

<sup>134</sup> TIM, Vol. XXIII, No. 11, Nov. 1956, p. 4.

The “republican country” that Tharchin referred to here is Egypt, and his reference was to the Suez Crisis. The Anglo–French–Israeli tripartite attack on Egypt over its nationalisation of the Suez Canal in the fall of 1956 resulted in condemnation by the international community, the exerting of diplomatic pressure on Britain, France, and Israel, and the first ever deployment of the United Nations Emergency Force, which was used to secure a cease-fire on the Egyptian–Israeli border.<sup>135</sup> However, as with his reference to the Hungarian Uprising, Tharchin’s comparison of the UN’s lack of response to the suppression of uprisings in eastern Tibet by the PLA to the UN’s intervention in the Suez Crisis is a tenuous one, since Egypt had been a UN member state since 1945.

In April 1959, three years after his reference to the Suez Crisis and one month after the 1959 Uprising in Lhasa, Tharchin turned again to the topic of the 1956 Hungarian Uprising. This time, his purpose was to show that he was not alone in seeing similarities between the Soviet suppression of Hungarian protesters and the suppression of uprisings in Tibet by the PLA:

According to a news report from March 30, [1959] from the Malay capital of Kuala Lumpur, the government of the Federation of Malaya stated: the attempt of the Chinese communist government to annihilate [protesters] in Tibet [by the means of] ruthless oppression is identical to [the case of] the 1956 Hungarian Uprising against the USSR.<sup>136</sup>

Behind this appeal to the concurrence of his views with those of the Malayan government, one can see Tharchin’s effort to convince his readers that the lack of peaceful settlement of the Tibet Question would lead to the instability in the whole world, as he had put it in his Suez article of 1956.<sup>137</sup> For Tharchin, the underlying aim was thus to impress upon his readers the broader importance of Tibet’s geopolitical position.

---

<sup>135</sup> See the UN General Assembly Resolution 1001 (ES-I) adopted on Nov. 7, 1956 (Available online at “Resolution 1001 (ES-I)” *United Nations Digital Library*. <https://digitallibrary.un.org/record/208418>, accessed July 24, 2024).

<sup>136</sup> TIM, Vol. XXV, No. 11, Apr. 1959, p. 8.

<sup>137</sup> *gal srid bod du bde ba ma byung nal 'dzam gling yongs su bde ba zhig e yong.*

### 3 Conclusion

The articles cited in this paper comprise only a fraction of the discourses and narratives found in *The Tibet Mirror* during the 1950s and 1960s. However, even this limited selection of examples shows how, in constructing a metanarrative of Tibetan independence, Tharchin did not only create vivid images of a Tibetan Self and a Chinese communist Other;<sup>138</sup> they show that he also aimed to construct an image of international support for Tibet in its struggle against the PRC. The main members of this support group were, as he saw it, Great Britain, India, the US, and the Kuomintang government. In this narrative, the Kuomintang was presented as, in contrast to the CCP, consistently supportive of Tibetan Buddhism and of Tibetan immigrants to Taiwan, as well as in close contact with the powerful “mother of Asian countries,” that is, the US. The *Tibet Mirror* also depicted the Kuomintang government as having resolutely guaranteed that it would grant independence to Tibet should it regain power, a claim not found in other historical sources.

Furthermore, Tharchin brought to the attention of his readership such essential aspects of the struggle for independence as the need to internationalise the Tibet Question and to engage with modern international organisations, such as the UN. The editor demonstrated how the Tibet Question should be presented to the international community, which facts and historical parallels could be used in arguing for Tibetan independence, and where to seek support for the Tibet Question.

The ideas and arguments that Tharchin laid out in *The Tibet Mirror* were not without influence. Many, if not all, of them would find their way into the policy and advocacy documents subsequently prepared by the Tibetan government-in-exile and exile politicians. In the booklet “Tibet Proving Truth from Facts,” released by the Department of Information and International Relations of the Central Tibetan Administration in India in 1996, the arguments in support of the claim

---

<sup>138</sup> For more details regarding the images of the Other and the Self, please refer to Moskaleva 2020.

for Tibetan independence cited the evidence Nepal had presented of its “independent diplomatic relations with Tibet” in its application for UN membership in 1949, just as Tharchin had noted at the time.<sup>139</sup> The same booklet referred to the speech on the Tibet Question delivered by Nehru in the Lok Sabha in 1959 as proof of India’s recognition of “Tibet’s right to self-determination,” which Tharchin had also done in *The Tibet Mirror* at the time (*Tibet Proving Truth from Facts*: 14). Similarly, in 2005, the CTA compiled a 300-page publication entitled “International Resolutions and Recognition of Tibet (1959 to 2004),”<sup>140</sup> containing all formal references to Tibet made by international bodies and legislatures. The exceptional role of India and the United States as the “two most important supporters” (Gyari 2022: 626) of the CTA as well as the “strong support base” (Gyari 2022: 429) of the Tibetan cause by the British people and individual royalty has been highlighted and thoroughly discussed by, for example, the 14<sup>th</sup> Dalai Lama’s Special Envoy in the US Lodi Gyaltsen Gyari (Rgya ri Blo gros rgyal mtshan, 1949–2018). Tharchin’s efforts to design and promote a discursive strategy for the pursuit of Tibetan independence from the end of the 1940s until 1963, when he published the last issue of *The Tibet Mirror*, prefigured what would become the dominant mode by supporters of that cause over the following half-century and beyond.

---

<sup>139</sup> “Nepal not only concluded peace treaties with Tibet and maintained an Ambassador in Lhasa, but also formally stated to the United Nations in 1949, as part of its application for UN membership, that it maintained independent diplomatic relations with Tibet” (*Tibet Proving Truth from Facts*: 5).

<sup>140</sup> Lobsang Nyandak Zayul, Kalon. *International Resolutions and Recognition of Tibet (1959 to 2004)*. Department of Information and International Relations, Central Tibetan Administration. Dharamsala, India. n.d. [2005].

## Bibliography

Brox, Trine

“Tibetan Culture as Battlefield: How the Term ‘Tibetan Culture’ is Utilized as Political Strategy.” In L. Schmithausen (ed.) *Buddhismus in Geschichte und Gegenwart: Gewalt und Gewaltlosigkeit im Buddhismus*, Vol. X, Hamburg: Universität Hamburg, 2006, pp. 85–105.

Claude, Richard P.

“Book Review. The International Commission of Jurists: Global Advocates for Human Rights,” *Human Rights Quarterly* 16 (3), 1994, pp. 576–578. Available online at <https://www.jstor.org/stable/762438> (accessed January 19, 2025).

Danilova, Anna

*Manipulirovanie Slovom v Sredstvakh Massovoi Informatsii. 3-e izd.* [Verbal Manipulation in Mass Media. Third Edition]. Moscow: Dobrosvet and Izdatelstvo KDU, 2014.

Dreyfus, George

“Are We Prisoners of Shangrila? Orientalism, Nationalism, and the Study of Tibet,” *Journal of the International Association of Tibetan Studies* 1, 2005, pp. 1–21. Available online at [https://www.researchgate.net/publication/268361871\\_Are\\_We\\_Prisoners\\_of\\_Shangrila\\_Orientalism\\_Nationalism\\_and\\_the\\_Study\\_of\\_Tibet](https://www.researchgate.net/publication/268361871_Are_We_Prisoners_of_Shangrila_Orientalism_Nationalism_and_the_Study_of_Tibet) (accessed January 19, 2025).

Dvornichenko, Andrei, Yurii Tot, and Mikhail Khodyakov

*Istoriya Rossii: Uchebnik.* [History of Russia: Textbook]. Moscow: TK Velbi and Izdatelstvo Prospekt, 2008.

Fader, Louis H.

*Called from Obscurity: The Life and Times of a True Son of Tibet Gergan Dorje Tharchin.* Vol. 3. Kalimpong: Tibet Mirror Press, 2009.

Goldstein, Melvyn C.

*A History of Modern Tibet, Vol. 3. The Storm Clouds Descend: 1955–1957.*  
Berkeley: University of California Press.

Gyari Lodi Gyaltsen (Rgya ri Blo gros rgyal mtshan, 1949–2018)

*The Dalai Lama's Special Envoy: Memoirs of a Lifetime in Pursuit of a Reunited Tibet.* New York: Columbia University Press, 2022.

Knaus, John K.

"Official Policies and Covert Programs: The U.S. State Department, the CIA, and the Tibetan Resistance," *Journal of Cold War Studies* 5 (3), 2003, pp. 54–79. [doi:10.1162/152039703322286773](https://doi.org/10.1162/152039703322286773).

McKay, Alex

"Tibet and the British Raj, 1904–47: The Influence of the Indian Political Department Officers." PhD Thesis. London: SOAS London University, 1995.

"'Kicking the Buddha's Head': India, Tibet and Footballing Colonialism." In P. Dimeo and J. Mills (eds.) *Soccer in South Asia: Empire, Nation, Diaspora*. London: Frank Cass, 2001, pp. 89–104.

"The British Invasion of Tibet, 1903–04," *Inner Asia* 14 (1), 2012, pp. 5–25. Available online at <https://www.jstor.org/stable/24572145> (accessed January 19, 2025).

Moskaleva, Natalia

"Sketches of Contemporary Tibetan History in *The Tibet Mirror* (1949–1963)," *Revue d'Etudes Tibétaines* (37), 2016, pp. 247–261. Available online at [https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret\\_37\\_14.pdf](https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_37_14.pdf) (accessed January 19, 2025).

"'What Does Babu Say?', a Pinch of Artistic Approach to News Reporting in *The Tibet Mirror* (1949–1963)," *Revue d'Etudes Tibétaines* (46), 2018, pp. 98–148. Available online at [https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret\\_55\\_18.pdf](https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_55_18.pdf) (accessed January 19, 2025).

“*The Tibet Mirror* and History Spinning in the 1950s and 1960s,” *Revue d’Etudes Tibétaines* (55), 2020, pp. 409–439. Available online at [https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret\\_55\\_18.pdf](https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_55_18.pdf) (accessed January 19, 2025).

Okawa, Kensaku

“Lessons from Tibetans in Taiwan: Their History, Current Situation, and Relationship with Taiwanese Nationalism.” 东洋文化研究所纪要. 第152册 [*Summary of the Institute of Research of East Asian Culture.*, 152], 2007, pp. 581–601. Available online at <https://repository.dl.itc.u-tokyo.ac.jp/record/26971/files/ioc152014.pdf> (accessed January 20, 2025).

Powers, John and David Templeman

*Historical Dictionary of Tibet*. Lanham: The Scarecrow Press, 2012.

“Question of Tibet: Resolution / Adopted by the General Assembly 1959,” *United Nations Digital Library*. Available online at <https://digitallibrary.un.org/record/664377?ln=en> (accessed July 24, 2024).

Rgyal lo don grub

*The Noodle Maker of Kalimpong. My Untold Story of the Struggle for Tibet*. Edited by Anne F. Thurston. London: Random House. 2015.

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.” PhD Diss. Heidelberg: Heidelberg University, 2018. [doi:10.11588/heidok.00025156](https://doi.org/10.11588/heidok.00025156).

Shakya, Tsering

*The Dragon in the Land of Snows: A History of Modern Tibet since 1947*. New York: Penguin Compass, 2000.

Singh, Teshu

“India, China and the Nathu La: Understanding Beijing’s Larger Strategy towards the Region,” *Institute of Peace and Conflict Studies*

(IPCS) *Issue Brief* (204), 2013, pp. 1–4. Available online at [https://www.ipcs.org/issue\\_select.php?recNo=491](https://www.ipcs.org/issue_select.php?recNo=491) (accessed January 19, 2025).

Smith, Paul J.

“Bhutan-China Border Disputes.” In B. Elleman, S. Kotkin, and C. Schofield (eds.) *Beijing's Power and Border Disputes: Twenty Neighbors in Asia*. Armonk: M.E. Sharpe, 2013, pp. 23–35.

Tenzin, Tashi C.

“Barmiok Athing Tashi Dadul Densapa: Colossus of Sikkim,” *Talk Sikkim* (7), 2010, pp. 37–39. Available online at <https://ru.scribd.com/document/35644325/Barmiok-Athing-Tashi-Dadul-Densapa-1> (accessed January 19, 2025).

*Tibet Proving Truth from Facts*. Dharamsala: Department of Information and International Relations, Central Tibetan Administration, n.d. [1996]. Available online at <https://tibet.net/tibet-proving-truth-from-facts-1996/> (accessed January 19, 2025).

Tikhvinskiy, Sergei (ed.)

*Istoriya Kitaya s Drevneishikh Vremen do Nachala XXI veka: v 10 t.* [History of China from Ancient Times to the Beginning of the 21<sup>st</sup> Century: 10 vols.], *Tom VIII. Kitaiskaya Narodnaya Respublika (1949–1976)* [Vol.8: The People's Republic of China (1949–1976)], edited by Yuri M. Galenovich. Moscow: Nauka, 2017.

TIM = *The Tibet Mirror* = *Yul phyogs so so'i gsar 'gyur me long* [The Mirror of News from Various Countries]. Kālimpong: edited by Dorje Tharchin, 1925–<1963>. Electronic reproduction. v. II, No. 5 (1927), v. V, No. 1 (1930)–v. XXVIII (1963). New York: Columbia University Libraries, 2009. Available online at [http://www.columbia.edu/cu/lweb/digital/collections/cul/texts/ldpd\\_6981643\\_000/](http://www.columbia.edu/cu/lweb/digital/collections/cul/texts/ldpd_6981643_000/) (accessed July 24, 2024)

Tolley, Howard B. Jr.

*The International Commission of Jurists: Global Advocates for Human Rights*. Pennsylvania: The University of Pennsylvania Press, 1994.



“UN General Assembly Resolution 1001 (ES-I) adopted on November 7, 1956,” *United Nations Digital Library*. Available online at <https://digital.library.un.org/record/208418> (accessed July 24, 2024).

Yurlov, Felix and Evgenia Yurlova

*Istoriya Indii*. [History of India. 20<sup>th</sup> Century]. Moscow: Institute of Oriental Studies RAS, 2010.

Zayul, Lobsang Nyandak

*International Resolutions and Recognition of Tibet (1959 to 2004)*. Dharamsala: Department of Information and International Relations, Central Tibetan Administration, n.d. [2005].

## Appendices

### *Appendix 1*

Tharchin’s account of Nehru’s statement on Tibet to the Lok Sabha, published in *the Tibet Mirror*, Vol. XXV, No. 11, Apr. 1959, pp. 2–3.

I [would like to] briefly report on a few key points from a detailed discussion of Tibet by the esteemed Prime Minister [J.] Nehru during a meeting of the Indian Council of Ministers [*sic*] Lok Sabha<sup>141</sup> on March 30, [1959] in New Delhi.

It has been about three years since the Indian government welcomed [here] the independent autonomous government of Tibet that was in the shadow of China. When the Indian government gained its complete independence from British rule, other regions that had previously been taken under British control did not think that India would take control [into its own hands] and refused to accept this seizure of power.

In terms of politics, by acting in accordance with the law, India cannot do anything about the Tibet [Question]. However, from a religious point of view, based on its close, longstanding cultural and

---

<sup>141</sup> *rgya gar log sa bā zhes pa’i bka’ shag lhan rgyas*. Lok Sabha (लोकसभा, lit. “House of the People”) is the lower house of India’s bicameral Parliament.

religious ties with Tibet, [India] feels profound sympathy for Tibetans and [sees] the damage which has been brought by communist China to the numerous monastic communities as a result of its unprincipled violence and which has become the reason for the [total] decay of good virtues [there].

Since we want to maintain close ties and pure mutual friendship with Tibetans, we urge [them] to expand their freedom and independence. At the same time, it is extremely important for us to maintain [our] friendly relations with China as well. Although it [China] is such a state, [we] do not capitulate to them [i.e., the Chinese] [when they] give an order to India.

Although India and Tibetans have strong kinship, cultural, and other ties, we have not interfered in the boiling [process] of Tibetan politics. Therefore, when earlier, about 55 years ago [1904], the Indian government sent troops into Tibet under the command of the officer [Francis] Younghusband, although [we] certainly intervened, it was not a daunting intervention with a takeover sanctioned by the government's order. It is a legacy that has been inherited [by us] from Great Britain and the Indian government of the time.

When seizing control over the Tibetan territories, India was always renouncing [its right to] them so that [they] would keep [the same status as before]. However, no matter what policy regarding the Tibet [Question] has been pursued [by India], in response communist China has established the practice of not following [a similar path]. We had no desire to establish such a rule in any country of the world. [...]

Although every Chinese government claimed that Tibet belonged to China, successive Tibetan governments did not recognise [it]. We should not interfere with the existing laws, or reality, and [should] act as witnesses.

The esteemed [Prime Minister] Nehru recalled that when the Premier of the People's Republic of China Zhou Enlai arrived in India two and a half years ago [in 1956], the two of them discussed the situation in Tibet. He [Nehru] will not now recall what questions [they] asked each other with regard to that situation. However, [Nehru said the following:] "According to his [Zhou Enlai's] statement then, although Tibet was part of China, Tibet was not China. Since Tibet had always been a subordinate territory of China with its [own] autonomous government, [China] even had the intention to grant full autonomy [to Tibet]. As I [Nehru] remember, he made such statements

about politics concerning the Tibet Question. I myself was very much pleased that he insisted on granting Tibet full autonomy and independence. I replied to him [Zhou Enlai]: ‘If, in accordance with what you said, Tibet is granted full autonomy and independence, the difficulties that arose earlier in Tibet will surely diminish.’”

## *Appendix 2*

Tharchin’s description of the aid provided by India, Sikkim and Kalimpong, published in TIM, Vol. XXVI, No. 1, June 1959, suppl. 1:

The Indian government has provided asylum for all those [people] who went into exile from Tibet after the Dalai Lama. For about 10,000 monks and laymen who came through Tawang [Rta dbang] in the Mon region and a few thousand monks and laymen who came through Bhutan—over 13,000 people in total—the Indian government has made arrangements for their temporary settlement in [the area] called Mussoorie and in [the area] called Buxa Dooars at the border with Bhutan. [The Indian government] continues to improve [their living] conditions [there].

Because of the efforts to relocate [Tibetans] to cooler areas within a short period of time, the government of Sikkim—by virtue of its enormous mercy to Tibetans—also provided over 2,000 people with work on road construction in the cooler areas of Sikkim, and by now, over about 1,000 [Tibetans] have relocated to Gangtok.

The esteemed government of Sikkim provides significant assistance in treating many [Tibetans] who fell ill because of the hot [weather] conditions. Moreover, during a brief encounter in Gangtok, I witnessed how the honourable Mrs. Phunkhang (Phun khang)<sup>142</sup> of the [11<sup>th</sup>] Dalai Lama’s<sup>143</sup> family together with the Princess of Sikkim and the son of the distinguished Sikkimese official ‘Bar thing<sup>144</sup> examined and comforted the sick on a daily basis, for which I must express my gratitude.

---

<sup>142</sup> *yab gzhi phun khang lha lcam mchog*.

<sup>143</sup> Mkhas grub rgya mtsho (1838–1856).

<sup>144</sup> ‘bar thing. Tharchin probably referred to a son of the Sikkimese official Barmiok Athing (1904–1988). See Tenzin Tashi 2010.

Besides, many lamas and *tulku* set off on their own and have arrived in Kalimpong because the heat made it uncomfortable for them to stay in Mussoorie. However, although virtually everyone may experience difficulties, what matters most is that the venerable Dalai Lama and the Tibetan government will certainly make every possible effort [for the good of Tibetans].

Apart from that, the Kalimpong district is arranging the necessary conditions for collecting donations for about 100 refugees. The esteemed Acharya Kripalani<sup>145</sup> has recently arrived here. When [he] petitioned for the appointment of the Assembly of Administrative Managers, the head of administrative managers replied that help would come. However, conditions were not yet right, but there was much hope that [conditions] would be right soon.

The Indian government also pays considerable attention and makes arrangements for more than 1,500 scholar monks with a *geshe* degree to get relocated to the Buxa area and [start] teaching [there], and [makes] arrangements for more than 500 lamas who are lineage holders [of the Buddhist teachings] to arrive and settle in Punjab in the area called Dalhousie, as well as makes every possible effort for ordinary monks.

### *Appendix 3*

Summary of K. M. Munshi's speech on Tibet, published in TIM, Vol. XXVI, No. 6–7, Nov.–Dec. 1959, p. 9.

According to the latest news from Agra from December 13, [1959], the leader of a group called "Independent India," who had previously served as the governor of Uttar Pradesh, Mr. K.M. Munshi, held a two-day meeting on the Tibet Question. This is a rough translation of what he said then:

"We heard that now in Tibet there is a terrifying situation with the oppression [of people] with unbearable suffering. For this reason, we have sincerely been greatly distressed too. Given that earlier Tibet used to be an independent religious state just like us, the introduction of any malevolent and highly improper measures [aimed] at the restriction of human rights and the religious sphere of an independent state are

---

<sup>145</sup> Acharya Kripalani (1888–1982) was an Indian politician and independence activist.

nothing but indecent evil acts which violate ethical norms. Moreover, by combining an ancient political strategy with communist ideology, communist China has been exacerbating the terrible situation in Tibet.

Since this does not concern solely Tibet and is spreading at our doorway, we cannot [afford] to make the slightest mistake. Those people in China called the Han have conducted and are conducting a campaign of violent expansionism. They have said and are saying that they will liberate not only Tibet, [but] all people of Asia.

As was stated by esteemed [Finance minister] Shakabpa [Zhwa sgab pa], the minister and political frontman (the mastermind of the political strategy) whom the venerable 14<sup>th</sup> Dalai Lama sent [here] specifically for this meeting, communist China has unbearably oppressed and is [currently] oppressing Tibet. [Chinese communists] have undertaken the extermination of the fine ancient traditional religion and culture of the entire Tibetan nation, as well as the social norms, rights, and basic welfare of [all] human beings [there]. Some people believe that, except only for a few lamas, monks, and officials, [Chinese communists] do not treat in such an inappropriate way [any] other [people]. Although [some people] think so, this is not true. Communist China has been destroying the entire Tibetan nation regardless of the noble or low origin [of people], of their religion, or of anything else. Therefore, it is important to help [lay] the foundation of peace in Tibet with a fair ethical-legal system.”

#### *Appendix 4*

“Fears That the Happiness of Minority Nationalities of the Himalayas Is Going to Be Devoured by Red Demons from the East,”<sup>146</sup> published in TIM, Vol. XXVII, No. 6, Feb.–Mar. 1961, p.6.

Nowadays, communist China is obsessed with the prosperity and power of the independent Indian state, and it is applying various harsh and gentle strategies to satisfy its hunger.

If one asks why [communist China uses] such strategies? [It does so] to swallow and bully the key areas which are similar to fingers in the

---

<sup>146</sup> *hi ma la ya'i ri bsul du chags pa'i mi rigs grangs nyung rnams kyi bde skyid la shar phyogs srin dmar gyis za sems la dwogs zon.*

north of India—the regions of upper Ladakh and the lower eastern Assam, [as well as such] neighbouring states [as] Nepal, Sikkim, and Bhutan in between them. On a pretext of feigned love, [Chinese communists] are trying to spread the communist strategies [there] and isolate [these areas] from India. In other words, [Chinese communists] are trying to harm the strength of India so that it becomes like, for example, a palm without fingers which is left without any strength.

Besides, earlier, for example, there was an independent Tibetan Religious State between India and China, [and] therefore the two powerful states did not need to meet face to face and they were able to exist [each] in one's own place. Or [one can say that] the independent Tibetan state stood like a border guard for the two mutually prosperous states.

However, when communist China invaded Tibet in 1950, even though the independent Tibetan government submitted a petition to the United Nations and to the government of [its] neighbour India, since [the petition] was disregarded and neglected, there was not anyone who paid attention to [it] and gave it [any] credibility.

Furthermore, since Tibet was perceived as a constituent of China, the independent Tibetan state has no choice but to remain under the rule of communist China at present. As all people, high and low, monks and laymen from Kham and Tibet did not foresee [the danger], they have all suffered from being fooled by the deceitful words of Chinese communist demons in the 17-Point Agreement, which is like a razor with smeared honey and a hat made of wet leather. As for now, look at the real experience of suffering in the human realm from committing a suicide with the razor, from squeezing one's own head by the hat made of wet leather, and from other hardships.

Similarly, now Chinese communists will be telling pleasant demonic lies to the small states and peoples [located] between India and Tibet. It will be very good if [you] stay very cautious regarding that and do not get fooled. Otherwise, [you] will [also] inevitably [experience] what [you] see and hear [now] about the situation in China and Tibet.

With the 17-Point Agreement

Having been fooled, the entire Tibetan nation has been annihilated.

With the Five Principles of Peaceful Coexistence

Having been fooled, India has been put in an uneasy position.

*Appendix 5*

Tharchin’s “rough translation” of Nepal’s application for membership of the UN, published in TIM, Vol. XVIII, No. 1, Oct. 1949, p. 5.

According to a recent news report, the Nepalese government has filed a petition to join the United Nations. A rough translation from English of the Nepalese government’s response to the questions of the United Nations is presented below:

“The Nepalese government is an independent state. [We] were able to protect our lands on our own. In addition to the fact that we send our official representatives to foreign countries, [we] are not under the rule of any powerful country. Earlier in 1815, the Gorkha [Kingdom], or Nepal, was at war with the British, and there was even a peace treaty signed. Also, when [we] attacked China in 1792 and when we fought with Tibet in 1855, at the time when the treaties were concluded, [we] made decisions independently without any third party or an intermediary state. Considering [all the above-mentioned facts] and the fact that [certain] internal adjustments were made [in the country], Nepal is requesting permission to join the United Nations.

The claim that the property and income of Nepal are controlled by India is not true. Since Nepalese-owned resources and goods are comparable to Indian ones, [we] develop [our] trade relations with foreign countries. In order to be able to defend the territory of Nepal on our own, [we] have well-equipped troops and sufficient weaponry for conducting military operations. There is not a single foreigner in the Nepalese government and military structures. The Nepalese government can even declare a war and order the mobilisation of its troops. [It also] has the power to sign a peace treaty on its own.”

*Appendix 6*

Tharchin’s appeal to Khampas to rise up. TIM, Vol. XXI, No. 4, July 1953, p. 3.

Brothers residing in the Tibetan region of Kham! Due to the need conditioned by the current circumstances, you are [all] brothers [to each other]. Although it is probably impossible to disobey the orders

of the Other [Chinese communists], rise up in the near future for the independence like Tito!

The Tito[ists] residing in the Tibetan region of Kham! In order for your motherland to remain an independent state forever, [you,] being creative, enthusiastic, and very brave, rise up in the near future for the independence like Tito!

I, the old man, know that there are many Tito[ists] in the Tibetan region of Kham. Enthusiastic and very brave Tito[ists]! Rise up in the near future for independence like Tito!

