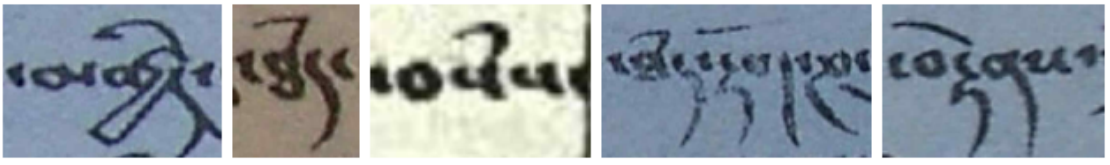
Transkribus in Practice: Abbreviations

In my **last article (https://digitalorientalist.com/2022/10/25/transkribus-in-practice-improving-cer/)**, I wrote about my experiments to improve the accuracy of a Handwritten Text Recognition (HTR) model using ***Transkribus* (https://readcoop.eu/de/transkribus/)**. This is part of my ongoing work with the ERC-funded project, ***The Dawn of Tibetan Buddhist Scholasticism (11th-13th C.)* (https://www.oeaw.at/ikga/tibschol)** (TibSchol),[1] at the Austrian Academy of Sciences. The goal is to see if *Transkribus* can train Handwritten Text Recognition (HTR) model(s) that can automatically process Tibetan cursive (*dbu med*) manuscripts of works from the 11th to 13th centuries, making a large amount of the early bKa' gdams pa (བཀའ་གདམས་པ་) scholastic corpus text searchable. When checking the transcripts in our training data, I noticed that our manuscripts contained a variety of abbreviations: some of these had been written out in full, while others transcribed exactly what they read in the manuscript, for example, *laso* or *la sogs* (ལ་སོགས་; and so forth). To improve the accuracy of our model, we had to consider how we wanted the abbreviations to be reproduced and to standardise the characters used.

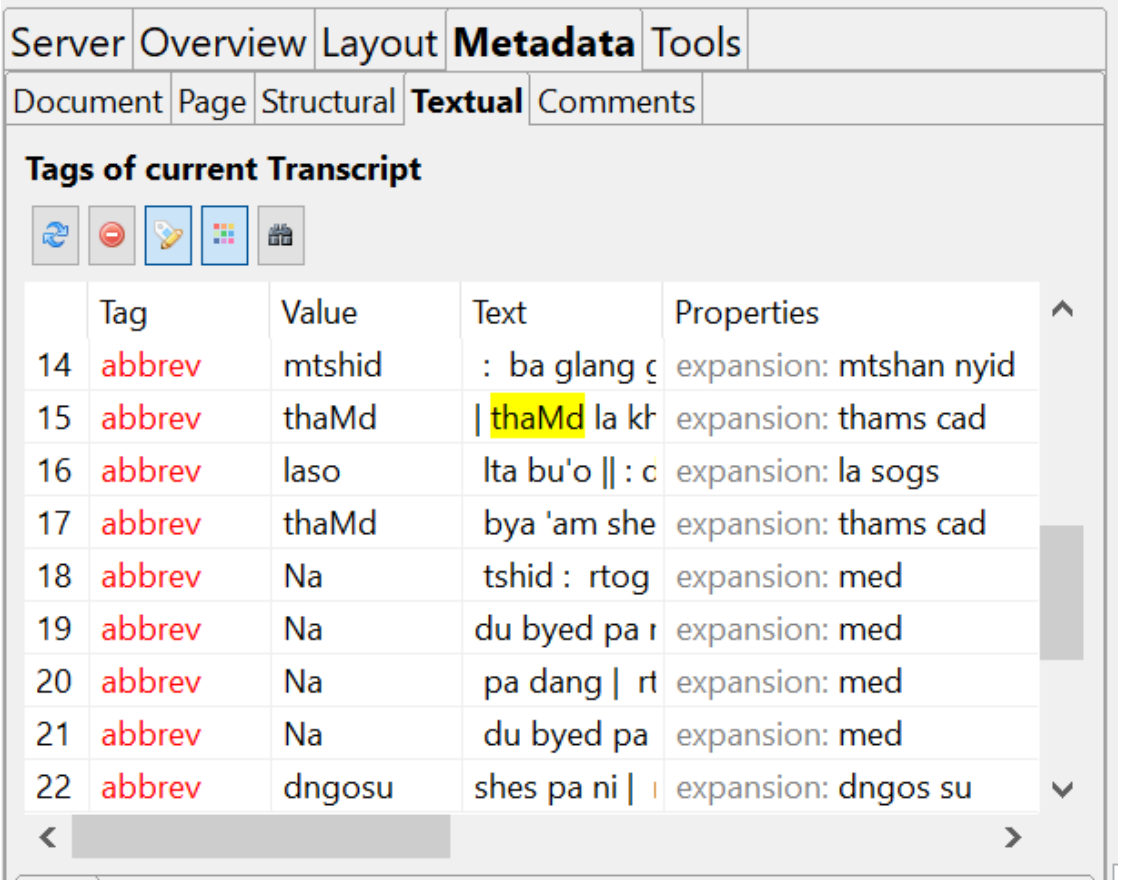
The first step was deciding how we wanted to deal with abbreviations in *Transkribus*; should they be reproduced as it appears in the manuscript, or should they be written out in full? While less faithful to what is represented on the manuscript, the latter can make text searching easier, especially due to the variety of abbreviations for a single word. For example, in just one manuscript *mtshan nyid* (མཚན་ཉིད་; nature, characteristic) was abbreviated in four different ways; *mtshyid* (མཚྱིད་), *mtshid* (མཚིད་), *tshyid* (ཚྱིད་), and *tshid* (ཚིད་). And so, searching for the full expansion, in this case *mtshan nyid*, will yield more accurate results than searching for *mtshyid*. **Models trained in Latin (https://readcoop.eu/acta_17-greifswald-supercharged/)** have shown that simple and frequently used abbreviations, such as *est*, *et*, and *per*, can be successfully learned by a HTR model when taught consistently. Of the 48 abbreviations identified in our material so far, only four of these appear more than 10 times in our training data, with most occurring between two and five times. This low rate of occurrence suggests it would prove challenging for our model to recognise and accurately transcribe these in full. Moreover, most of the abbreviations we have found are contractions, which can cause further difficulties for HTR.



(https://digitalorientalist.files.wordpress.com/2022/10/abbreviations.png)
Examples of abbreviations identified

Based on the frequency and complexity of the abbreviations identified, it was decided they would be transcribed as they appeared in the manuscript. Most abbreviations consist of a group of letters taken from the full version of the word/phrase. The manuscripts also contain a few abbreviation characters, which are reproduced using additional characters, for example, *thaMd* for *thams cad* (ཐམས་ཅད་; all, entire, whole). Each abbreviation and its transcription are recorded in a spreadsheet for reference in future

parts of the project. I then tag the abbreviations in the "Textual" tab, within the "Metadata" tab, and include the expansion of the abbreviation (see screenshot below). The expansion of the abbreviation then becomes part of the metadata, which can be exported in TEI and DOCX formats (along with the abbreviation). Moreover, tags can be included in model training (for both PyLaia HTR and CITlab HTR+), which includes the training of abbreviations with their expansions. I have tried this and found that the trained model was able to identify and tag some common Tibetan abbreviations, such as *thaMd*. However, the expansion is usually included in the transcript itself instead of the "Metadata" tab. So the transcription would read *thaMdthams cad*, instead of *thaMd*.



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

Screenshot of "Textual" tab with a list of abbreviations tagged in a single page

I believe this approach offers the most possibilities for engaging with the material produced; a text search of the expansion is still possible whilst also recording the abbreviated form for further investigations in the future, such as palaeographic or codicological analysis. However, this approach is laborious because every abbreviation must be tagged. As I mentioned in my previous post, the TibSchol project started with existing transcriptions, which means reading all our initial transcripts, amending abbreviated words, and then tagging them. There is also the possibility of finding-and-replacing abbreviated words using the search function of *Transkribus*, although this can be awkward when navigating large pages of results. This approach would likely work best for those working with a small pool of abbreviations.

There are alternatives to marking up abbreviations by hand. The ***Bentham Project*** **(https://readcoop.eu/mastering-latin-abbreviations-and-hyphenations-the-bentham-and-deeds-projects/)**, for example, which has transcribed almost 25,000 pages of English philosopher Jeremy Bentham's writings, has created an abbreviation dictionary, which connects directly to *Transkribus* via an API script coded by Ismail Prada. The script uses its find-and-replace algorithm to locate terms found in the abbreviation dictionary, replace them with its shorter equivalent, and tag them as abbreviations. Compiling a dictionary of abbreviations is, in itself, time intensive. This is true of Tibetan, at least, where there are currently limited resources on Tibetan abbreviations. As such, this process strikes me as one that would be of more benefit to those working with large amounts of material, like the Bentham Project, and/or those whose documents are heavily abbreviated.

When I started experimenting with abbreviations, I noticed that there was little information on how others have approached this in their training. I hope this opens up more discussion, and if you have tips for working with abbreviations, I would love to hear them!

Endnotes

[1] This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001002 TibSchol). The results presented are solely within the author's responsibility and do not necessarily reflect the opinion of the European Research Council or the European Commission who must not be held responsible for either contents or their further use.

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

Cover Image: Phywa pa chos kyi sengge. *dGe tshul gyi tshig le'ur byas pa sum brgya pa'i tshig don rab tu 'byed pa*. Par gzhi dang po'i par thengs dang po. 1 vols. Gangs can khyad nor dpe tshogs. Lha sa: Ser gtsug nang bstan dpe rnying 'tshol bsdu phyogs bsgrigs khang, 2019. Accessed November 1, 2022. **http://purl.bdrc.io/resource/W3CN22740 (http://purl.bdrc.io/resource/W3CN22740)**. [BDRC bdr:W3CN22740]

Published by

Rachael Griffiths

View all posts by Rachael Griffiths

NOVEMBER 1, 2022NOVEMBER 1, 2022 LEAVE A COMMENT HTR, TIBETAN STUDIES, TRANSCRIBING, TRANSKRIBUS