Journal of Open Humanities Data

Text Recognition for Nepalese Manuscripts in Pracalit Script

DATA PAPER

ALEXANDER JAMES O'NEILL
NATHAN HILL
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This dataset is a model for handwritten text recognition (HTR) of Sanskrit and Newar Nepalese manuscripts in Pracalit script. This paper introduces the state of the field in Newar literature, Newar manuscripts, and HTR engines. It explains our methodology for developing the requisite ground truth consisting of manuscript images and corresponding transcriptions, training our model with a PyLAia engine, and this model's limitations. This dataset shared on Zenodo can be used by anyone working with manuscripts in Pracalit script, which will benefit the fields of Indology and Newar studies, as well as historical and linguistic analysis.

CORRESPONDING AUTHOR:
Alexander James O'Neill
Department of East Asian Languages and Cultures, SOAS University of London, London, UK
ao34@soas.ac.uk

KEYWORDS:
handwritten text recognition; PyLAia; Transkribus; Sanskrit; Newar; Manuscripts

TO CITE THIS ARTICLE:
O'Neill, A. J., & Hill, N. (2022). Text Recognition for Nepalese Manuscripts in Pracalit Script. *Journal of Open Humanities Data*, 8: 26, pp. 1–6. DOI: https://doi.org/10.5334/johd.90

(1) OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.6967421.

CONTEXT

Newar (also referred to as Nepāl Bhāṣā) is the indigenous language of the Kathmandu Valley. In its pre-print phase, this highly literate and creative culture produced thousands of works that have remained mainly unstudied in either western or Nepalese scholarship. Much of Newar literature is a mixture of Newar, Sanskrit, and Maithili (Malla, 1981, 6–9). While Newar literature is written in various scripts, the most common by far is the Pracalit script, which has thus also come to be known as Newar Lipi (Newar script) (Pandey, 2012). Thus, for both Indological interest in Nepalese manuscripts written in Sanskrit and for students of Newar language and culture, a means to compile a digital corpus more quickly through optical character recognition (OCR) becomes apparent.

OCR engines have gradually become more effective in recent decades. Handwritten text recognition (HTR) has proven to be far more problematic. Deep learning neural networks have made it possible to build HTR models based on images of handwritten text linked with corresponding transcriptions (called "ground truth"). A character error rate (CER) under 10% allows for effective automatic transcription (Muehlberger et al., 2019). Advances in computing power and storage made by the Transkribus platform developed by READ-COOP have enabled the training of large data sets involving multiple hands, allowing for generalised HTR models for particular writing styles (Hodel et al., 2021). Transkribus hosts two HTR engines: CITIlab-HTR+ (Michael et al., 2018) and PyLaia, a PyTorch-based model (Mocholí Calvo et al., 2018).

In principle, models for HTR of Indic texts can be developed similarly to those in Roman scripts. Transkribus already has two publicly available HTR+ models for printed 19th and 20th century Devanagari developed by Nicole Merkel-Hilf (2022). This project focused on expanding the abilities of HTR models to Indic texts in pre-print and non-Devanagari sources, focusing on Sanskrit and Newar (Nepāl Bhāṣā) manuscripts in Pracalit script from the 16th to 19th centuries.

(2) METHOD

An HTR trainer requires diplomatic transcriptions of Pracalit manuscripts to line up with text in manuscript photographs. Critically edited editions can speed up transcription and ground truth generation through de-correction. Databases like GRETIL, from which we sourced the published transcriptions, make it possible to bootstrap a non-existent HTR model by using texts from other scripts (Georg-August-Universität Göttingen, 2020). To this end, transcriptions were prepared based on the following four Nepalese manuscripts, each with different varieties of Pracalit script. For each entry in the list below, in order, the manuscript title is given in italics followed by call numbers in parentheses, deposit location, manuscript languages and date, and sources of the corresponding transcriptions:

1. *Hitopadeśa* (MIK I 4851)
   Staatsbibliothek zu Berlin
   Mixed Newar and Sanskrit, 1561 CE
   Original transcription by Alexander James O'Neill
2. *Vetālapañcaviṃśati* (HS. Or. 6414)
   Staatsbibliothek zu Berlin
   Newar, 1675 CE
   Adapted transcription based on unpublished materials by Felix Otter (Otter, n.d.a)
3. *Avalokiteśvaraguṇakāraṇḍavyūha* (MS Add. 1322)
   Cambridge Digital Library
   Sanskrit, 18th century
   Adapted transcription based on an edition by Lokesh Chandra (Chandra, 1999)
4. *Madhyamasvayaṃbhūpurāṇa* (RAS Hodgson MS 23)
   Royal Asiatic Society Online Collection
   Mixed Newar and Sanskrit, c. 1800

Adapted transcription based on unpublished materials by Felix Otter (Otter, n.d.b) and the published Nagarjuna Institute transcription (Shakya & Bajracharya, 2001)

While the HTR+ engine appeared to have difficulty working with the lack of word division, PyLaia produced better results, and we used it for the rest of the training. We trained the model on 441 pages of manual transcriptions of the above four manuscripts, with validation performed on 242 pages that were not part of the training set. It was further tested and continues to be used on pages that were not part of the training or validation sets. We decided it would be most appropriate and culturally sensitive to transcribe into Unicode Pracilt (Unicode, Inc., 2021), see Figure 1.

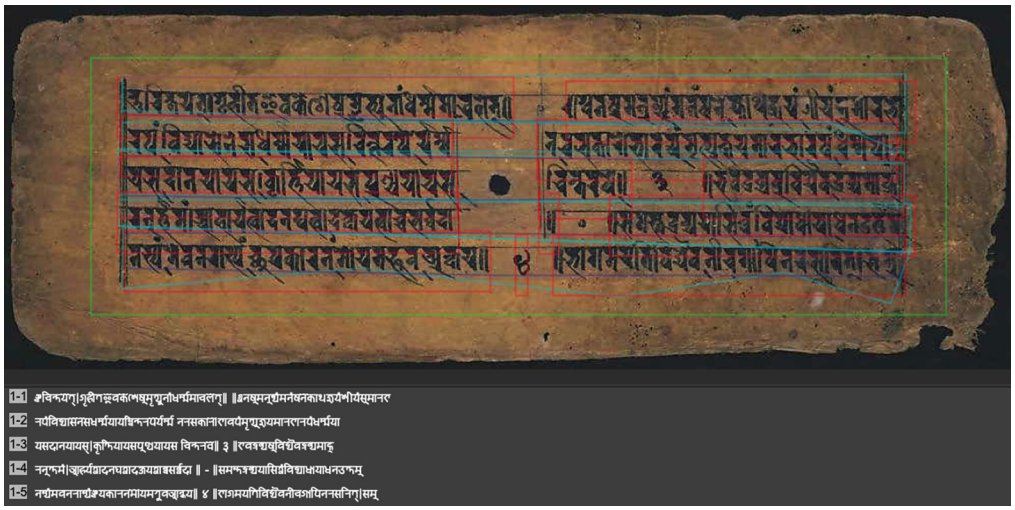

Figure 1 Screenshot of a completed transcription of a folio of *Hitopadeśa* (MIK I 4851) in Transkribus.

Using 250 epochs, Transkribus trained a model with a CER on the training set of 2.6% and 0.1% on the validation set. This discrepancy may signify little more than that the latter had fewer complex characters to recognise. Therefore, the model produces accurate results when transcribing the same or similar hands to those responsible for these four manuscripts, see Figure 2.

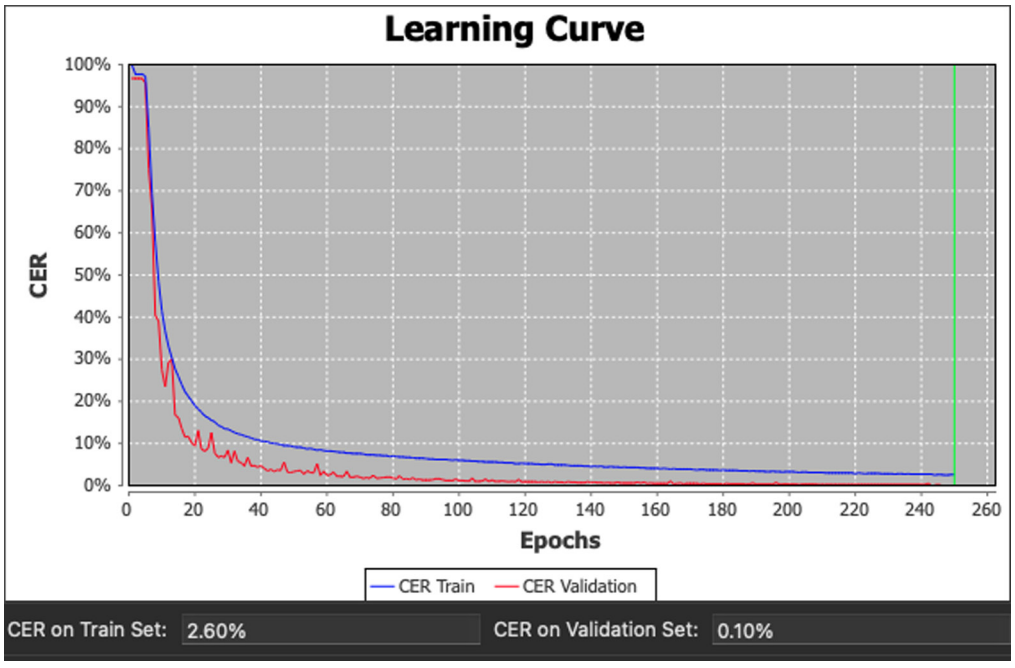

Figure 2 Screenshot of the model's learning curve on Transkribus.

QUALITY CONTROL

The model has a higher CER when applied to irregular forms of Pracalit script, including more ornate or rougher hands (Figure 3) However, with a trained base model, new hands require

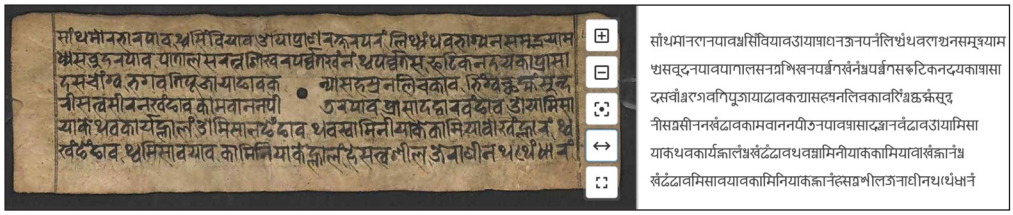

Figure 3 An example of a cruder form of Pracalit, from *Vetālapañcaviṃśati* (HS. Or. 6414), transcribed on Transkribus.

significantly fewer pages, ranging from ten to thirty pages of new ground truth. We will update and refine the model with new ground truth as we encounter variant hands.

The main limitation of this model's initial and continued training is the lack of transcriptions. However, bootstrapping existing editions and transcriptions and feeding corrected machine-generated transcriptions back into the model are workable solutions.

In transcription, the model encounters difficulties with damaged or soiled manuscripts, irregular spacing, punctuation, and illustrations interrupting the text. It is worth noting that while the vast majority of Pracalit manuscripts are written in a *scriptio continua*, occasional spacing and irregular punctuation conventions produce mixed results for the model. While mistakes in ground truth produce incorrect transcriptions, a larger mass of correct ground truth reduces the impact of any one mistake.

(3) DATASET DESCRIPTION

Object name – OCR model for Pracalit for Sanskrit and Newar MSS 16th to 19th C., Ground Truth

Format names and versions – png and xml

Creation dates – 2022-04-01 – 2022-08-04

Dataset creators – Alexander James O'Neill, SOAS University of London, Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization

Language – Sanskrit and Newar

License – Creative Commons Attribution 4.0 International

Repository name – Zenodo

Publication date – 2022-08-05

(4) REUSE POTENTIAL

While it is possible to share models within Transkribus, this has limited potential for the shared creation of ground truth. As modelled by the GitHub collection "HTR united," which combines the ground truth of French documents (Chaqué & Clérice, 2021), it is possible to make ground truth data sets available in ways that others can use within platforms such as Transkribus and elsewhere. We have therefore made our dataset publicly available on Zenodo in the form of PNG and XML files that can be used on HTR platforms (O'Neill, 2022). For the future, in collaboration with the Centre of Asian and Transcultural Studies (CATS) Bibliothek at the University of Heidelberg, we are participating in the development of a South Asian Studies-specific ground truth database in a FID4SA (Fachinformationsdienst für Südasien: Specialised Information Service for South Asia) dataverse, called "Ground truth data for HTR on South Asian Scripts," as part of the University of Heidelberg's research data archive heiDATA (Universität Heidelberg, 2022).

As the most labour-intensive part of philological practice, the ability to quickly produce machine-readable transcriptions of various witnesses of an Indic text is of great value to Indology and other disciplines. This enables high-speed searches and comparisons of corpora, as well as linguistic analysis through machine-learning methods (Meelen et al., 2021). In disciplines such as Newar studies, where there is both a paucity of trained scholars and a profusion of manuscripts, this tool can contribute to easing the burden of compiling and editing a digital corpus, which will benefit linguistic, literary, and historical analysis of the Newar language by easing the burden of work with primary manuscript sources.

ACKNOWLEDGEMENTS

We would like to extend our thanks to Felix Otter (Philipps-Universität Marburg) for providing us with transcriptions.

FUNDING INFORMATION

This work was funded by the Arts and Humanities Research Council (AHRC), UKRI, as part of the project "The Emergence of Egophoricity: a diachronic investigation into the marking of the conscious self." Project Reference: AH/V011235/1. Principal Investigator: Nathan Hill, SOAS University of London.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Alexander James O'Neill: Data curation, Formal Analysis Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Nathan Hill: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

AUTHOR AFFILIATIONS

Alexander James O'Neill orcid.org/0000-0001-9982-2589
Department of East Asian Languages and Cultures, SOAS University of London, London, UK
Nathan Hill orcid.org/0000-0001-6423-017X
Department of East Asian Languages and Cultures, SOAS University of London, London, UK; Trinity Centre for Asian Studies, Trinity College Dublin, Dublin, Ireland

REFERENCES

Chandra, L. (Ed.) (1999). *Guṇakāraṇḍavyūhasūtram*. International Academy of Indian Culture.

Chaqué, A., & **Clérice, T.** (2021). *HTR-United*. GitHub. https://github.com/HTR-United/htr-united (last accessed: 9 November 2022).

Georg-August-Universität Göttingen. (2020). *GRETIL: Göttingen Register of Electronic Texts in Indian Languages and related Indological materials from Central and Southeast Asia*. GRETIL. Retrieved from http://gretil.sub.uni-goettingen.de/gretil.html (last accessed: 22 August 2022).

Hodel, T., Schoch, D., Schneider, C., & **Purcell, J.** (2021). General Models for Handwritten Text Recognition: Feasibility and State-of-the-Art. German Kurrent as an Example. *Journal of Open Humanities Data*, *7*(13), 1–10. DOI: https://doi.org/10.5334/johd.46

Malla, K. P. (1981). *Classical Newari Literature*. Nepal Study Centre.

Meelen, M., Roux, E., & **Hill, N.** (2021). "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*, *20*(1), 1–11. DOI: https://doi.org/10.1145/3409488

Merkel-Hilf, N. (2022). Ground Truth data for printed Devanagari [Dataset]. In *FID4SA@heiDATA*. DOI: https://doi.org/10.11588/data/EGOKEI

Michael, J., Weidemann, M., & **Labahn, R.** (2018). *HTR Engine Based on NNs P3: Optimizing speed and performance - HTR+* [Deliverable 7.9 for READ project funded by EU Horizon 2020 Project 674943]. READ-COOP. Retrieved from https://readcoop.eu/wp-content/uploads/2018/12/Del_D7_9.pdf (last accessed: 8 November 2022).

Mocholí Calvo, C., Vidal Ruiz, E., & **Puigcerver i Pérez, J.** (2018). *Development and experimentation of a deep learning system for convolutional and recurrent neural networks* [Degree final work]. Universitat Politècnica de València. Retrieved from https://riunet.upv.es/bitstream/handle/10251/107062/MOCHOL%C3%8D%20-%20Desarrollo%20y%20experimentaci%C3%B3n%20de%20un%20sistema%20de%20aprendizaje%20profundo%20para%20redes%20neuronale....pdf?sequence=1&isAllowed=y (last accessed: 8 November 2022).

Muehlberger, G., Seawrd, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Culluto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., … Zagoris, K. (2019). Transforming scholarship in the archives

through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. DOI: https://doi.org/10.1108/JD-07-2018-0114

O'Neill, A. (2022). OCR model for Pracalit for Sanskrit and Newar MSS 16th to 19th C., Ground Truth [Dataset]. In *Zenodo*. DOI: https://doi.org/10.5281/zenodo.6967421

Otter, F. (n.d.a). *Vetālapañcaviṃśati* [Unpublished transcription].

Otter, F. (n.d.b). *Madhyamasvayaṃbhūpurāṇa* [Unpublished transcription].

Pandey, A. (2012). *Proposal to Encode the Newar Script in ISO/IEC 10646* [Proposal from the Script Encoding Initiative]. eScholarship. https://escholarship.org/uc/item/50c8w93x

Shakya, M. B., & **Bajracharya, S. H.** (Eds.) (2001). *Svayambhū Purāṇa*. Nagarjuna Institute of Exact Methods.

Unicode, Inc. (2021). *Newa Range: 11400–1147F* [Excepted Character Code tables for The Unicode Standard, Version 14.0]. Unicode. Retrieved from https://www.unicode.org/charts/PDF/U11400.pdf (last accessed: 8 November 2022).

Universität Heidelberg. (2022). *Ground truth data for HTR on South Asian Scripts*. FID4SA@heidata. Retrieved from https://heidata.uni-heidelberg.de/dataverse/FID4SA-GT (last accessed: 9 November 2022).

TO CITE THIS ARTICLE:
O'Neill, A. J., & Hill, N. (2022). Text Recognition for Nepalese Manuscripts in Pracalit Script. *Journal of Open Humanities Data*, 8: 26, pp. 1–6. DOI: https://doi.org/10.5334/johd.90

Published: 30 November 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.