

Shin Buddhist Comprehensive **Research Institute Buddhist Manuscript Research Project**



CENTRE FOR ASIAN AND TRANSCULTURAL STUDIES

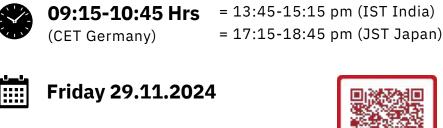
SOUTH ASIA INSTITUTE Department of Cultural and Religious History of South Asia



Joint Research Group Manuscriptology and Digital Humanities

Text Recognition for South Asian Scripts – Using Transkribus

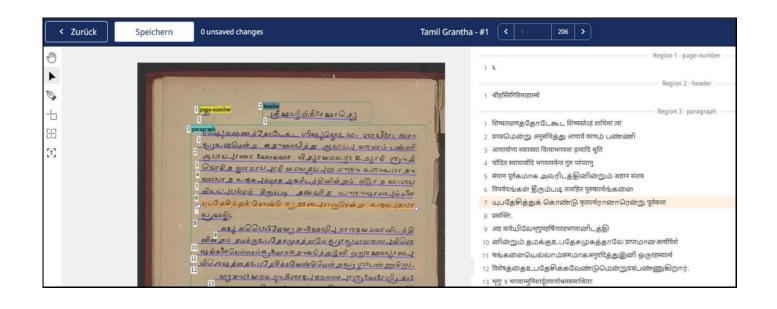
@heiDATA	•		
	HTR on South Asian Scripts armation Service South Asia)		
heiDATA > FID4SA@heiD	ATA > Ground truth data for HTR on South Asian Scripts >		
Ground Ti	ruth data for printed Devanagari		
	Merkel-Hilf, Nicole, 2022, "Ground Truth data for printed Devanagari", <u>https://doi.org/10.11588/data/EGOK</u> EL heiDATA, V1 Cite Dataset - Learn about <u>Data Citation Standards</u> .	Access Dataset +	
wages, survey, she		Contact Owner Dataset Metrics 😧 727 Downloads 😡	Share
Description	Ground truth (GT) data (jpg and alto xml files) for an OCR model that recognizes printed text in Devanagari script.	TET DUMINAUS	
	The GT data was trained on Transkribus with the HTR+ engine. The training was performed on appr. 220 pages with appr. 27,000 words. The validation set was 10% of the training set.		
	The training material is comprised of letterpress printings from the Naval Kishore Press (Lakhnau, North India) from the late 19th and early 20th century in the Hindi, Sanskrit, Braj Bhasha and Awadh linguages.		
	Transcription was performed by Nicole Merkel-Hilf (CATS Library / Heidelberg University Library) with support by Daria Peshcherova (CATS Library / Heidelberg University Library). (2022-10-24)		
Subject 😧	Arts and Humanities		
Keyword 9	OCR (Optische Zeichenerkennung), Optical Character Recognition, Text recognition, Ground Truth, Devanagari, Devanagari Character Set (Data processing)		
License/Data Use	CC BY 4.0		





ZOOM LINK

Speaker: Nicole Merkel-Hilf and Sowndarya Sriraman (Universität Heidelberg)



AI-based approaches to text recognition have taken text recognition models with Transkribus for printed

OCR for historical documents a big step forward, especially for documents written in non-Latin scripts. Since 2019, the Naval Kishore Press – digital project has been using the Transkribus platform for the entire recognition process (layout analysis, text recognition, post-processing) for the Devanagari texts in the collection. The project "Hindu Temple" Legends in South India" conducted at the Heidelberg Academy of Sciences and Humanities has trained

Tamil / Grantha texts.

In the first part of the talk Nicole Merkel-Hilf will present the workflow and results from the NKP project by using the Transkribus Expert Client. In the second part of the talk Sowndarya Sriraman will present the workflow and results for Tamil and she will also give a live presentation on how to use the Transkribus' web application.

Department of Cultural and Religious History of South Asia (Classical Indology) Heidelberg University • Centre for Asian and Transcultural Studies • South Asia Institute Voßstrasse 2, Building 4130 • 69115 Heidelberg • www.sai.uni-heidelberg.de

