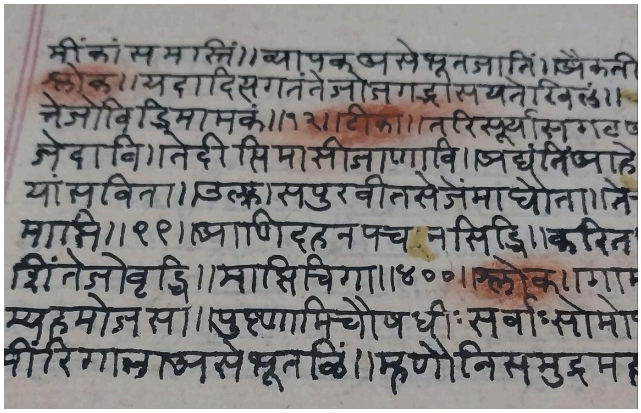
Democratizing Wisdom: A Journey Through Time with Dnyaneshwari

By Gauri Bhagwat

Student - Master Rare Book and Digital Humanities, UBFC, France

DHBenelux 2024 Conference submission for consideration for a short paper

Bhagavad Gita, part of the Mahabharata- the Sanskrit epics of ancient India, is a 700-verse Hindu scripture, dated to the second half of the first millennium BCE. Understanding Sanskrit was challenging for the common people. Recognizing this, Sant Dnyaneshwar composed a commentary in vernacular Marathi the Dnyaneshwari - in 1290 AD. This profound text played a pivotal role in bringing the philosophical teachings of the Gita, which were previously the prerogative of Sanskrit-knowing elitists, to a larger public.



Dnyaneshwari is made of 9,032 verses, in which each verse consists of four lines called an Ovi (ōvī). Dnyaneshwari's detailed explanations made it comprehensible for all. For instance, a single Gita verse on the characteristics of a knowledgeable person is expanded in 328 verses in Dnyaneshwari, while the topic of ignorance is elaborated in 248 verses. In total, the 700 Gita shlokas are explained in 9,032 verses using numerous examples and similes to enhance understanding. This approach made spiritual pursuit accessible and shaped attitudes towards life

for a diverse audience. Thus, Dnyaneshwari remains highly relevant even today for individuals interested in spiritual exploration and life philosophy.

The 1584 AD version of Dnyaneshwari by Sant Eknath remains a key source for further commentaries and interpretations. There are differences in these commentaries based on the editor and version. For example during the medieval period commentary was written with a focus on the society which had different social norms, values and attitude than versions prepared in the 20th century. Hence there is a need for a comparative study through the centuries that extends beyond linguistic nuances, offering a profound understanding of the intricate interplay between language, culture, and society. In order to investigate this, this research seeks to develop a Text Recognition AI model for transcribing Devanagari script, encompassing handwritten - printed texts followed by a comparative study on transcriptions to analyze the changes observed in the Dnyaneshwari over the past 300 years.

To achieve this I consider **a corpus of the source manuscript** of Dnyaneshwari, 1751 AD, Paper; and **three printed Dnyaneshwari** (20th Century) Editions prepared by the following editors: Rajwade, Dandekar, and Godbole.
Various versions and editions of Dnyaneshwari were searched in libraries, Institutes and private collectors. The one of the oldest manuscripts was found in The Bhandarkar Oriental Research Institute, Pune, India. From them the digital copies of the required chapters were obtained. I found the older and more authentic editions of the printed books with private collectors and with their due permission the photos of the chapters were taken.

These digital copies were then cleaned, cropped and HD optimized for better OCR/HTR results. I decided to try out the Transkribus platform for getting OCR/HTR results. Transkribus is an AI-powered platform used for the text recognition, image analysis and structure recognition of historical documents. There are public and private text recognition models on their platform. Unfortunately there is no public model for the Devanagari manuscripts on the platform, so I created a model myself. As I have both handwritten and printed material, **I have created two models. A HTR model for the 18th - 19th century and an OCR model for the 20th century.** Both are trained for the Devanagari script containing Marathi, Prakrut Marathi, Hindi and Sanskrit languages.

Training for a Devanagari script posed unique challenges due to the script's intricate nature. Specifically Devanagari comprises 34 consonants and 14 vowels.

अ	a	आ	ā
इ	i	ई	ī
उ	u	ऊ	ū
ए	e	ऐ	ai
ओ	o	औ	au
अं	ṃ	अः	ḥ

Vowels

क ka	ख kha	ग ga	घ gha	ङ ṅa
च ca	छ cha	ज ja	झ jha	ञ ña
ट ṭa	ठ ṭha	ड ḍa	ढ ḍha	ण ṇa
त ta	थ tha	द da	ध dha	न na
प pa	फ pha	ब ba	भ bha	म ma
य ya	र ra	ल la	व va	
श śa	ष ṣa	स sa	ह ha	

Consonants

Every Vowel has an independent form, but when combined with a consonant it forms a diacritic form as follows:

	a	**ā**	**i**	**ī**	**u**	**ū**	**ē**	**ai**	**ō**	**au**	**ṁ**	**ḥ**	**ê**	**ô**
Vowels	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	ऍ	ऑ
	ka	**kā**	**ki**	**kī**	**ku**	**kū**	**kē**	**kai**	**kō**	**kau**	**kaṁ**	**kaḥ**	**kê**	**kô**
Consonant with Diacritics	क	का	कि	की	कु	कू	के	कै	को	कौ	कं	कः	कॅ	कॉ

The combination of 34 consonants in its diacritics form presents a complex matrix with 476 variations. There are also conjunctions of consonants resulting in ligatures like त्+ व = त्व (tva). There are multiple consonant ligatures like ग्न्य (gny), त्स्त्र (tstr), क्ष्ण्य (kṣṇy) and even a word

त्स्न्र्य (rtsnya) with 5 consonants as a single ligature. The most variations in conjunctions can be seen for the consonant र (Ra). For example:

र (ra) +	ट (ṭa)	ट्र (ṭra)
	श (śha)	श्र (śhra)
	त (ta)	त्र (tra)
	व (va)	र्व (rva)

Devanagari is the fourth most widely adopted writing system in the world, but as most of the OCR softwares are Anglocentric, there is just one model available on Transkribus. The existing model is trained on printed books from a specific press, thus limiting the OCR accuracy for other sources. Also in the case of manuscripts there were variations in writing styles of conjunctions, numbers and sometimes even normal letters, making it more challenging for text recognition. After careful and meticulous training, two models were created on Transkribus. One for the Devanagari manuscripts (18th - 19th Century) and other for the printed books (20th century).
In this presentation **I will present this process,two models and the HTR & OCR challenges for this domain. In addition, I would like to get feedback from the community for the next phase of the project; namely how NLP can be implemented to compare these versions.** Using NLP Python Libraries such as iNLTK (Natural Language Toolkit for Indic Languages) and Indic NLP Library, the process of word tokenization will be initiated, which will then be used for Word to word comparison andsentence comparison.
This can be further used to add syllable distinguisher and transliteration for those who have difficulty reading Devanagari script.

In conclusion, this research strives to bridge literary gaps by creating a HTR/OCR AI model. Dnyaneshwari serves as a poignant example of historical evidence of democratizing knowledge. By delving into the evolution of this pivotal Marathi text over centuries, the research contributes not only to academic discourse but also to the broader societal narrative of embracing diversity and breaking silos for the greater good.

Glossary (Pronunciation):
Bhagavad Gita - buhg-uh-vuh-d gee-tah (bhagavadgītā)
Devanagari - dē-vɐ-na-gɐri
Dnyaneshwari - ɲyɐ-a-nē-sh-war-i (D is silent)
Sant is not equivalent to but close to English translation Saint
Sant Dnyaneshwar - Sɐnt ɲyɐ-a-nē-sh-war (D is silent)
Sant Eknath - Sɐnt Ēh-k-nah-th

Bibliography:

Bansal, V., & Sinha, R. M. K. (2001). A complete OCR for printed Hindi text in Devanagari script. *Proceedings of Sixth International Conference on Document Analysis and Recognition*, 800–804. https://doi.org/10.1109/icdar.2001.953898

Davis, R. H. (2015). *The Bhagavad Gita: A biography*. Princeton university press.

Gahankari, A., Kapse, A. S., Atique, M., Thakare, V. M., & Kapse, A. S. (2023). Hybrid approach for word sense disambiguation in Marathi language. *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*. https://doi.org/10.1109/gcat59970.2023.10353377

Gupta, V., Joshi, N., & Mathur, I. (2019). Advanced machine learning techniques in Natural Language Processing for Indian languages. *Smart Techniques for a Smarter Planet*, 117–144. https://doi.org/10.1007/978-3-030-03131-2_7

Indira, B., Shuaib Qureshi, M., Sharief Shaik, M., Mahmood Saqib, R., & V Ramana Murthy, M. (2012). Devanagari character recognition: A short review. *International Journal of Computer Applications*, *59*(6), 23–27. https://doi.org/10.5120/9553-4011

Kale, S., & Gawande, U. (2021). Implementation of automatic mesh rules generation by word features relationship (AMRG-WFR) for word sense disambiguation for Marathi language. *International Journal of Next-Generation Computing*. https://doi.org/10.47164/ijngc.v12i5.456

Khandale, K. B., & Mahender, C. N. (2020). Natural language processing based rule based discourse analysis of Marathi text. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. https://doi.org/10.1109/icesc48915.2020.9155653

Patil, A., & Dwivedi, P. (2024). Enhanced recognition of handwritten Marathi compound characters using CNN-SVM hybrid approach. *Fusion: Practice and Applications*, *14*(2), 26–42. https://doi.org/10.54216/fpa.140202

Sengupta, P., & Chaudhuri, B. B. (1993). Natural language processing in an Indian language (bengali)-i: Verb phrase analysis. *IETE Technical Review*, *10*(1), 27–41. https://doi.org/10.1080/02564602.1993.11437284

Shukla, A., Sharma, A., Aggarwal, A., & Jain, S. (2023). Devnagari character recognition using Optical Character Recognition (OCR). *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. https://doi.org/10.1109/cictn57981.2023.10141358

Yadav, D., Sanchez-Cuadrado, S., & Morato, J. (2013). Optical character recognition for Hindi language using a neural-network approach. *Journal of Information Processing Systems*, *9*(1), 117–140. https://doi.org/10.3745/jips.2013.9.1.117