The proceedings of the Computational Sanskrit and Digital Humanities section

under the

19th World Sanskrit Conference

Kathmandu, Nepal

June 26 – 30, 2025

Edited by

AMBA KULKARNI & OLIVER HELLWIG

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-250-3

Preface

This volume presents peer-reviewed and edited versions of shortlisted papers accepted for presentation in the session Computational Sanskrit and Digital Humanities at the 19th World Sanskrit Conference, taking place from June 26-30 2025 in Kathmandu. We received a total of 24 submissions each of which was reviewed by at least three reviewers. This process resulted in 12 submissions shortlisted for presentation at Kathmandu. The accepted papers show that Sanskrit NLP continues to grow as a field, with increasing attention to literary analysis, manuscript digitization, tool development, and the integration of large language models (LLM) into the workflow. The papers can be grouped into four main areas: literary and poetic analysis, text digitization and recognition, structural and morphological modeling, and tools and resources.

A first group of papers deals with the analysis of Sanskrit poetry and literary features. **Jadhav et al.** present a method to classify *upamā alaṇkāra*s using large language models. The Llama-3.1 based system distinguishes between complete and elliptical similes and can also extract their components. **Sandhan et al.** propose a human-in-the-loop framework to study the qualities of fine Sanskrit poetry. Providing annotations of the *Śikṣāṣṭaka* from the views of different schools of Sanskrit poetics, the authors explore how *kāvyaśāstra* principles can be combined with computational techniques to accelerate data acquisition. **Kulkarni and Neelamana** examine itaretara dvandva compounds, which can express both conjunctive and disjunctive meanings. The paper explains why such compounds are hard to interpret, especially when nested within larger compound structures. Examples from classical Āyurveda texts show these difficulties in practice.

Another group of papers focuses on digitizing Sanskrit texts using technologies like OCR, handwritten text recognition, and automatic speech recognition (ASR). **Chincholikar et al.** present a pipeline to recognize handwritten Sanskrit manuscripts, with case studies from early modern Sanskrit literature. The system includes line segmentation, text recognition using Devanāgarī OCR, and post-correction using a Sanskrit LM. **Tsukagoshi et al.** address the problem of OCR for printed Vedic Sanskrit, which requires special attention to accent marks. Their system employs Transformer-based OCR models and converts Devanāgarī to ISO 15919 transliteration, supporting both philological research and dataset creation. **Kumar et al.** introduce *Vedavani*, the first large dataset and benchmark for ASR in Sanskrit Vedic poetry. The 54-hour dataset includes over 30,000 audio samples from the *Ṛg-* and *Atharvaveda*, along with evaluations of different multilingual speech models.

A third group of papers focuses on grammatical structure and semantic representation. **Krishnan et al.** propose a new method to analyze Sanskrit compounds, combining constituency and dependency analysis in a unified framework. Being based on Pāṇinian principles, their system includes rules and lexical information to help identify compound structure and types. **Lakkundi et al.** present a markup language, IKML (Indic Knowledge Markup Language), for encoding the structure of *śāstra*s. IKML allows scholars to annotate Sanskrit texts with grammatical, semantic, and discourse-level information. The paper also introduces a collaborative web interface, e-Bhāṣya, for annotation and visualization. The contribution by **Krishnan et al.** focuses on aligning *Ṛgveda-padapāṭha* annotations from three digital platforms, aiming to

create a unified Vedic database and develop a morphological analyzer for Vedic Sanskrit.

The final group of papers introduces tools and resources to improve access to Sanskrit texts and support digital scholarship. **Neill** presents *Pāṇḍitya*, a web-based tool to visualize relationships between Sanskrit authors and their works. It allows users to explore commentarial networks and connect to online Sanskrit e-texts. **Jagadeeshan et al.** study Sanskrit document retrieval using English queries, focusing on extracts from the *Bhāgavatapurāṇa*. The paper compares several retrieval strategies and shows that translation-based methods work best. In addition, the authors release a new data set of over 3,000 English-Sanskrit query-document pairs that can serve as ground truth for further developments. **Patel** explores how to build a concordance of Sanskrit synonyms using *samānārthaka kośa*s, laying the foundation for future lexicographic work.

We thank the Convenors, Programme Committee members and the numerous experts who helped us in the review process, and all our authors who responded positively to the reviewers comments and improved their manuscripts accordingly. We would also like to express our thanks to the 19th WSC organising committee, led by Prof. Kashinath Nyaupane.

Amba Kulkarni & Oliver Hellwig

Programme Committee

- **Convenors:**
  - Gérard Huet (Inria, Paris, France)
  - Amba Kulkarni (University of Hyderabad, India)
- **Chairs:**
  - Amba Kulkarni (University of Hyderabad, India)
  - Oliver Hellwig (University of Zurich, Germany)
- **Members:**
  - Tanuja Ajotikar (Sanskrit Library, USA)
  - Ivan Andrijanić (University of Zagreb)
  - Stefan Baums (University of Munich)
  - Arnab Bhattacharya (IIT Kanpur)
  - Brendan Gillon (McGill University)
  - Pawan Goyal (IIT Kharagpur)
  - Amrith Krishna (Learno.ai)
  - Malhar Kulkarni (IIT Bombay)
  - Philipp Mass (University of Leipzig, Germany)
  - Patrick McAllister (IKGA, Austria)
  - Dhaval Patel (Ahmedabad)
  - Jivnesh Sandhan (IIT, Dharwad)
  - Pavankumar Satuluri (IIT, Roorkee)
  - Peter Scharf (Sanskrit Library, USA)
  - Sai Susarla (Siddhanta Knowledge Foundation, Chennai)

Table of Contents

An introduction to computational identification and classification of Upamā alaṅkāra
Bhakti Jadhav, Himanshu Dutta, Shruti Kanitkar, Malhar Kulkarni and Pushpak Bhattacharyya
1

Aesthetics of Sanskrit Poetry from the Perspective of Computational Linguistics: A Case Study Analysis on Śikṣāṣṭaka
Jivnesh Sandhan, Amruta Barbadikar, Malay Maity, Pavankumar Satuluri, Tushar Sandhan, Ravi M. Gupta, Pawan Goyal and Laxmidhar Behera 15

Itaretara Dvandva: A challenge for Dependency Tree semantics
Amba Kulkarni and Vasudha Neelamana 37

A Case Study of Handwritten Text Recognition from Pre-Colonial era Sanskrit Manuscripts
Kartik Chincholikar, Shagun Dwivedi, Kaushik Gopalan and Tarinee Awasthi 52

Towards Accent-Aware Vedic Sanskrit Optical Character Recognition Based on Transformer Models
Yuzuki Tsukagoshi, Ryo Kuroiwa and Ikki Ohmukai 70

Vedavani: A Benchmark Corpus for ASR on Vedic Sanskrit Poetry
Sujeet Kumar, Pretam Ray, Abhinay Beerukuri and Shrey Kamoji 81

Compound Type Identification in Sanskrit
Sriram Krishnan, Pavankumar Satuluri, Amruta Barbadikar, T S Prasanna Venkatesh and Amba Kulkarni 90

IKML: A Markup Language for Collaborative Semantic Annotation of Indic Texts
Chaitanya S Lakkundi, Gopalakrishnan Rajaraman and Sai Rama Krishna Susarla 109

Challenges in Processing Vedic Sanskrit: Towards creating a normalized dataset for the Ṛgveda-saṃhitā
Sriram Krishnan, Sepuri Gayathri and Amba Kulkarni 131

Pāṇḍitya: Visualizing Sanskrit Intellectual Networks
Tyler Neill 148

Anveshana: A New Benchmark Dataset for Cross-Lingual Information Retrieval on English Queries and Sanskrit Documents
Manoj Balaji Jagadeeshan, Prince Raj and Pawan Goyal 161

Concordance of Sanskrit Synonyms
Dhaval Patel 181