Proceedings of the Computational Sanskrit & Digital Humanities

Selected papers presented
at the 18^{th} **World Sanskrit Conference**

(in online mode)
Canberra, Australia
January 9 – 13, 2023

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 978-1-950737-71-0

Preface

This volume presents edited versions of shortlisted papers accepted for presentation in the session 'Computational Sanskrit and Digital Humanities' at the 18^{th} World Sanskrit Conference during Jan 9-13, 2023 in online mode. The physical conference was delayed by two years due to the global pandemic situation. Thus there were two rounds of calls for submissions. In order to disseminate the research among the scholars, the convenors of the Computational Sanskrit and Digital Humanities decided to hold an online event after the first round of submission. We thank the organisers of the World Sanskrit Conference and in particular the International Association of Sanskrit Studies for giving us permission to hold this event online. The papers shortlisted in the first round were presented at the online event during Jan 11-12, 2022, as well.

We received a total of 29 submissions in the two rounds. Each submission was reviewed by at least three reviewers and 15 submissions were shortlisted for presentation.

A good number of papers focus on how machine learning approaches can be applied to Sanskrit texts, paying special attention to its status as a low resource language. The contribution of Krishna et al. is an in-depth study of data-driven parsing algorithms for Classical Sanskrit. It discusses various parsing architectures, the choice of input features and possible sources of parsing errors. Another aspect of data-driven NLP is covered in the contribution of Sandhan et al. who compare static and contextualized word embeddings on a range of semantic classification tasks and obtain remarkably good results given the small size of the digital Sanskrit corpus. As a foundation for data-driven methods, Krishnan et al. discuss endeavors towards merging existing annotated corpora of Classical Sanskrit in a common format. Building NLP resources is also the topic of the paper by Sarkar et al. who develop a simple, yet effective machine learning approach for pre-annotating an NER data set. While the papers mentioned so far use quantitative models for language analysis, Maity et al. discuss the problem of disambiguating the kāraka notion of oblique cases from a Pāṇinian perspective and evaluate their approach on a small sample of annotated sentences.

The remaining papers of this volume cover aspects of (Sanskrit) NLP that go beyond the identification and analysis of basic linguistic structures. Mahesh and Bhattacharya evaluate how well high-level textual features can be predicted with contextualized word embeddings. The contribution of Neill describes the open-source project Skrutable, a tool for identifying meters in Sanskrit texts, and sketches directions for future research in this field. Another paper focusing on meter identification is presented by Terdalkar and Bhattacharya whose system also includes a fuzzy search option. Ajotikar et al. propose an extension of the TEI standard for encoding various levels of information present in the Sanskrit commentarial literature and discuss which research questions can be addressed by applying XSLT templates to textual sources encoded in this way. Another contribution by Scharf et al. describes the TEI encoding of the Rāmopakhyāna and the Kramapha system used to interactively display its content in the web. Terdalkar et al. report about ongoing work on building knowledge graphs from Āyurvedic texts. Their paper elaborates on the annotation process and its prerequisites as well as on the design of an appropriate medical and botanical ontology. The problem of extracting and representing the knowledge structure of śāstra texts is also addressed in the paper by Susarla et al. who concentrate on sentence level features of scientific texts. Hellwig et al. describe an extension of Bloomfield's concordance of Vedic mantras and discuss its application to open issues in Vedic Studies.

Finally, two papers address aspects of Middle Indo-Aryan languages and the Buddhist literature composed in them. Harnsukworapanich et al. give an overview of D-Tipitaka, an online version of

the Dhammachai Tipiṭaka Edition with linked manuscript images and contextual information. Zigmond discusses how Pāli commentaries can be distinguished from the commented canonical texts by applying standard clustering algorithms to frequent words in these texts.

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 reviewer's comments and improved their manuscripts accordingly. We thank the entire 18^{th} WSC organising committee, led by Prof McComas Taylor, which provided us the necessary logistic support for the organisation of this section.

Amba Kulkarni & Oliver Hellwig

Programme Committee

- **Convenors:**
  - Gérard Huet (Inria, Paris, France) (Chair)
  - Amba Kulkarni (University of Hyderabad, India)
- **Chairs:**
  - Amba Kulkarni (University of Hyderabad, India)
  - Oliver Hellwig (University of Zurich, Germany)
- **Members:**
  - Ivan Andrijani (University of Zagreb,Croatia)
  - Stefan Baums (University of Munich, Germany)
  - Arnab Bhattacharya (IIT Kanpur, India)
  - Brendan Gillon (McGill University,Canada)
  - Pawan Goyal (IIT Kharagpur, India)
  - Malhar Kulkarni (IIT Bombay, India)
  - Dhaval Patel (Ahmedabad, India)
  - Wiebke Petersen (University of Düsseldorf, Germany)
  - Pavan Kumar Satuluri (Chinmaya Vishwavidyapeeth, Veliyanad, India)
  - Sai Susarla (MIT, Pune,India)
  - Peter Scharf (IIIT Hyderabad, India)
  - Srinivasa Varakhedi (KKSU, Ramtek,India)

Table of Contents