



LREC 2026

**8th Workshop on Indian Language Data: Resources
and Evaluation (WILDRE-8)**

Workshop Proceedings

Editors

Girish Nath Jha, Kalika Bali, Sobha L, Devendr Kumar

12 May, 2026

Proceedings of WILDRE-8

©ELRA Language Resources Association (ELRA), 2026

These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-37-1
EAN 9782493814890

Preface

WILDRE – the 8th Workshop on Indian Language Data: Resources and Evaluation is being organised on May 12th, 2026, under the LREC 2026 platform. India has a high degree of linguistic diversity and has seen concerted efforts by the Indian government and industry to develop language resources. ELRA Language Resources Association and its associate organisations have been very active and successful in addressing the challenges and opportunities related to language resource creation and evaluation. It is, therefore, a great opportunity for resource creators in Indian languages to showcase their work on this platform and to interact and learn from those involved in similar initiatives worldwide.

The broader objectives of the 8th WILDRE will be

- to map the status of Indian Language Resources
- to investigate challenges related to creating and sharing various levels of language resources
- to promote a dialogue between language resource developers and users
- to provide an opportunity for researchers from India to collaborate with researchers from other parts of the world

The call for papers received a good response from the Indian language technology community.

Workshop Organisers

Organizing Committee

- Girish Nath Jha, Professor, School of Sanskrit and Indic Studies, Jawaharlal Nehru University, India
- Kalika Bali, Microsoft Research India Lab, Bangalore
- Sobha L, AU-KBC, Anna University
- Devendr Kumar, MGAHU, Wardha

Table of Contents

<i>MetricalARGS: Studying Metrical Poetry with LLMs</i> Chalamalasetti Kranti and Sowmya Vajjala	1
<i>Semi-automatic Approach for Tamil Discourse Relation Annotation</i> Frances Yung, Enosh Peter Ponraj and Vera Demberg	14
<i>Konkani Daan: A Community-Driven Culturally Grounded Speech Corpus for Low-Resource ASR</i> Milind Shivolkar, Vaibhav Gawas and Jyoti Pawar	25
<i>Is Literal Annotation Enough? Building an Annotation Framework for Metonymic Named Entities in Marathi</i> Pratibha Dongare	33
<i>Bengali-English and Hindi-English Code Mixed Speech Data with Disfluencies</i> Anuran Mitra, Tapabrata Mondal, Anirvan Chakravarty and Sivaji Bandyopadhyay	39
<i>Konkani Wordnet Resources</i> Hanumant H. Redkar, Mahadev Gawas, Anjali Desai and Jyoti Pawar	49
<i>Development of Speech Corpus for Low-Resource Language- A case of Sanskrit</i> Devendr Kumar, Girish Nath Jha and Khalid Choukri	55
<i>The shabd portal - searchable lexical resources for Indian languages by Government of India</i> mercy lalrohluo hmar, Girish Nath Jha and DHANANJAY SINGH	61
<i>POS Tagging in Low-Resource Maithili language: Specific Challenges and Nuances</i> Shivani Priya, Shruti Jha, Urmila Jha, Girish Nath Jha, Deepali Tiwari and Dr. Jyoti Raj	67
<i>Preserving Civilisation Memory: A Digital Humanities Approach to The Ramayan</i> Shashank Tiwari and Girish Nath Jha	75
<i>Naamah: A Large Scale Synthetic Sanskrit NER Corpus via DBpedia Seeding and LLM Generation</i> Annarao Kulkarni and Akhil Rajeev P.	88
<i>IndEuph-170: Benchmarking Cultural Pragmatics through Euphemism Detection in Indian English</i> Debamita Samajdar	93
<i>Integrating Syntactic and Discourse Signals through Multi-Encoder Fusion in NMT for Low-Resource Indian Language Pairs</i> Sobha Lalitha Devi, Vijay Sundar Ram and Pattabhi RK Rao	98
<i>NE-LID: A Fast and Accurate Language Identification System for Northeast Indian Languages</i> Badal Nyalang	104
<i>Integrating Cultural Wisdom and Digital Technologies for Children's Moral and Emotional Development</i> Ms. Garima and Girish Nath Jha	109

Workshop Program

12th May 2026

09:00–09:10

Welcome by Workshop Chairs

09:10–09:20

Inaugural Address by Kavita Bhatia

09:20–10:00

Keynote Talk by Amitabh Nag, Bhashini

10:00–11:00

Oral Talks I

10:00–10:15

MetricalARGS: Studying Metrical Poetry with LLMs

Chalamalasetti Kranti and Sowmya Vajjala

10:15–10:30

Semi-automatic Approach for Tamil Discourse Relation Annotation

Frances Yung, Enosh Peter Ponraj and Vera Demberg

10:30–10:45

Konkani Daan: A Community-Driven Culturally Grounded Speech Corpus for Low-Resource ASR

Milind Shivolkar, Vaibhav Gawas and Jyoti Pawar

10:45–11:00

Is Literal Annotation Enough? Building an Annotation Framework for Metonymic Named Entities in Marathi

Pratibha Dongare

11:00–11:20

Coffee Break

11:20 –11:50

Oral Talks II

11:20–11:35

Bengali-English and Hindi-English Code Mixed Speech Data with Disfluencies

Anuran Mitra, Tapabrata Mondal, Anirvan Chakravarty and Sivaji Bandyopadhyay

11:35–11:50

Konkani Wordnet Resources

Hanumant H. Redkar, Mahadev Gawas, Anjali Desai and Jyoti Pawar

11:50–12:50

Poster Session

11:50–12:50

Development of Speech Corpus for Low-Resource Language- A case of Sanskrit

Devendr Kumar, Girish Nath Jha and Khalid Choukri

11:50–12:50

The shabd portal – searchable lexical resources for Indian languages by Government of India

mercy lalrohluo hmar, Girish Nath Jha and DHANANJAY SINGH

11:50–12:50	<i>POS Tagging in Low-Resource Maithili language: Specific Challenges and Nuances</i> Shivani Priya, Shruti Jha, Urmila Jha, Girish Nath Jha, Deepali Tiwari and Dr. Jyoti Raj
11:50–12:50	<i>Preserving Civilisation Memory: A Digital Humanities Approach to The Ramayan</i> Shashank Tiwari and Girish Nath Jha
11:50–12:50	<i>Naamah: A Large Scale Synthetic Sanskrit NER Corpus via DBpedia Seeding and LLM Generation</i> Annarao Kulkarni and Akhil Rajeev P
11:50–12:50	<i>IndEuph-170: Benchmarking Cultural Pragmatics through Euphemism Detection in Indian English</i> Debamita Samajdar
11:50–12:50	<i>Integrating Syntactic and Discourse Signals through Multi-Encoder Fusion in NMT for Low-Resource Indian Language Pairs</i> Sobha Lalitha Devi, Vijay Sundar Ram and Pattabhi RK Rao
11:50–12:50	<i>NE-LID: A Fast and Accurate Language Identification System for North-east Indian Languages</i> Badal Nyalang
11:50–12:50	<i>Integrating Cultural Wisdom and Digital Technologies for Children's Moral and Emotional Development</i> Ms. Garima and Girish Nath Jha
12:50–13:00	Closing by Sobha Lalitha Dev