



LREC 2026

**The Second workshop on Challenges in Processing
South Asian Languages (CHiPSAL2026)**

Workshop Proceedings

Editors

Kengatharaiyer Sarveswaran and Ashwini Vaidya

16 May 2026

Proceedings of the Second workshop on Challenges in Processing South Asian Languages
(CHiPSAL2026)

©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 978-2-493814-66-1

Message from the Organising Chairs

Welcome to the proceedings of CHiPSAL 2026, the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL), held as part of the Language Resources and Evaluation Conference (LREC 2026). Building on the success of the inaugural workshop at COLING 2025, CHiPSAL 2026 continues to provide a platform for addressing the unique challenges of processing South Asian languages and for fostering collaboration among researchers working in this space.

South Asia is one of the most linguistically diverse regions in the world, yet its languages remain significantly underrepresented in language technologies, particularly in the era of large language models. This workshop brings together contributions that address key challenges such as data scarcity, morphological richness, script diversity, code-mixing, and the need for culturally grounded evaluation. The accepted papers span a wide range of languages, including both widely spoken languages and extremely low-resource and endangered languages such as Burushaski, Limbu, and Nepal Bhasha (Newari), reflecting the breadth and depth of current research efforts in the region.

We received 42 submissions to the main track, of which 21 papers were accepted for presentation, resulting in an acceptance rate of approximately 50%. Following withdrawals, 19 papers are included in the final proceedings. All submissions underwent a rigorous peer-review process, with each paper reviewed by three programme committee members. In total, 55 programme committee members from academia and industry worldwide contributed to the review process. In addition, the shared task received 15 system description paper submissions, of which 12 were accepted for publication.

CHiPSAL 2026 featured regular papers and a shared task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes for Nepali, which attracted strong participation from the research community. The diversity of submissions and the range of approaches highlight both the progress made and the challenges that remain in developing inclusive and effective NLP systems for South Asian languages.

We would like to express our sincere gratitude to the programme committee members for their timely and rigorous reviews. We also thank all authors for their valuable contributions and the LREC 2026 workshop chairs and organising committee for their support in facilitating this workshop.

We congratulate all authors of accepted papers and shared task contributions. We hope that CHiPSAL 2026 serves as a valuable forum for exchanging ideas, building collaborations, and advancing research on South Asian language technologies. We also look forward to expanding this initiative through future editions and regional events to further strengthen the research community in this domain.

For more information and future communications, please visit the workshop website: <https://sites.google.com/view/chipsal/>.

Organising Chairs of CHiPSAL 2026

Kengatharaiyer Sarveswaran, University of Jaffna, Jaffna, Sri Lanka.

Ashwini Vaidya, Indian Institute of Technology Delhi, India.

Bal Krishna Bal, Kathmandu University, Kathmandu, Nepal.

Surendrabikram Thapa, Virginia Tech, USA.

Tafseer Ahmed Khan, Mohammad Ali Jinnah University, Karachi, Pakistan.

Organising Committee

Organising Chairs

Kengatharaiyer Sarveswaran, University of Jaffna, Sri Lanka
Ashwini Vaidya, Indian Institute of Technology Delhi, India
Bal Krishna Bal, Kathmandu University, Nepal
Surendrabikram Thapa, Virginia Tech, USA
Tafseer Ahmed Khan, Mohammad Ali Jinnah University, Pakistan

Program Committee Members (alphabetical order)

A. M. Abirami, Thiagarajar College of Engineering (TCE), India
Ahrane Mahaganapathy, University of Jaffna, Sri Lanka
Aloka Fernando, IIT Colombo, Sri Lanka
Amba Kulkarni, University of Hyderabad, India
Aman Shakya, Pulchowk Campus, Tribhuvan University, Nepal
Ann Sinthusha Anton Vijeevaraj, University of Vavuniya, Sri Lanka
Arjuna S. R., Manipal Academy of Higher Education (MAHE), Bengaluru, India
Arafat Ahsan, IIIT Hyderabad, India
Ashwini Vaidya, IIT Delhi, India
Bal Krishna Bal, Kathmandu University, Nepal
Balaram Prasain, Central Department of Linguistics, Tribhuvan University, Nepal
Basanta Joshi, Pulchowk Campus, Tribhuvan University, Nepal
Buddhika Karunarathne, University of Moratuwa, Sri Lanka
Eugene Y. A. Charles, University of Jaffna, Sri Lanka
Faisal Alvi, Habib University, Pakistan
Farah Adeeba, University of Engineering and Technology, Pakistan
Gihan Dias, University of Moratuwa, Sri Lanka
Hariram Veeramani, UCLA, USA
Hassan Sajjad, Dalhousie University, Canada
Kritesh Rauniyar, Macquarie University, Australia
Krishna Chalise, Central Department of Linguistics, Tribhuvan University, Nepal
Lakshmojee Koduru, KPMG LLP, USA
Lekhnath Sharma Pathak, Central Department of Linguistics, Tribhuvan University, Nepal
Mazhar Ali, Benazir Bhutto Shaheed University Lyari, Pakistan
Meesum Alam, Indiana University, USA
Menan Velayuthan, Utrecht University, Netherlands
M. S. V. P. J. Sathvik, University of Birmingham, United Kingdom
Muhammad Rafi, FAST-NUCES, Pakistan
Muralikrishna S. N., MIT-MAHE, India
Munief Tahir, Al-Khawarizmi Institute of Computer Science, Pakistan
Mutee Ur Rahman, ISRA University, Pakistan
Naveed Unjum, FAU Erlangen-Nürnberg, Germany
Nisansa de Silva, University of Moratuwa, Sri Lanka
Parameswari Krishnamurthy, IIIT Hyderabad, India
Poonam Bansal, IGDTUW, India
Prakash Poudyal, Kathmandu University, Nepal
Rajani Chulyadyo, Kathmandu University, Nepal
Randil Pushpananda, University of Colombo School of Computing, Sri Lanka
Ritesh Kumar, UnReal-TecE LLP
Sahar Rauf, Al-Khawarizmi Institute of Computer Science, Pakistan
Sajjad Haider, Institute of Business Administration, Pakistan

Sana Shams, Al-Khawarizmi Institute of Computer Science, Pakistan
Sanjukta Ghosh, IIT BHU, India
Shuvam Shiwakoti, Virginia Tech, USA
Siddhant Bikram Shah, Northeastern University, USA
Suganya Ramamoorthy, Vellore Institute of Technology, India
Supriya Khadka, Coventry University, United Kingdom
Surabhi Adhikari, Columbia University, USA
Surangika Ranathunga, Massey University, New Zealand
Surendrabikram Thapa, Virginia Tech, USA
Tafseer Ahmed, Mohammad Ali Jinnah University, Pakistan
Toqeer Ehsan, VTT Technical Research Centre of Finland
Usman Naseem, Macquarie University, Australia
Uthayasanker Thayasivam, University of Moratuwa, Sri Lanka
Wazir Ali, Quaid-e-Awam University of Engineering, Science and Technology, Pakistan

Volunteers (alphabetical order)

Ahrane Mahaganapathy, University of Jaffna, Sri Lanka
Ajintha Sivakulasingam, University of Jaffna, Sri Lanka
Nishanthini Kanthakumar, University of Jaffna, Sri Lanka

Table of Contents

<i>Findings of the Second Workshop on Challenges in Processing South Asian Languages (CHiP-SAL 2026)</i>	
Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Ashwini Vaidya, Tafseer Ahmed and Bal Krishna Bal	1
<i>Development of Burushaski Speech - English Text Translation Dataset</i>	
Tauqeer Saleem, Abdul Samad, Azkaa Nasir, Adina Adnan Mansoor, Fatima Faisal and Mahrukh Yousuf	9
<i>Lost the Negation or Lost in Negation</i>	
Vennela Bairi and Parameswari Krishnamurthy	18
<i>A Morphological Transducer for the Limbu Language</i>	
Avyaya Singh and Jonathan N. Washington	32
<i>Evaluating Large Language Models for Medical Named Entity Recognition in Urdu: A Benchmark Study</i>	
Bushra Nasim, Kinza Latif, Muhammad Zohair, Muhammad Hassan Asif and Zarmeen Nasim	40
<i>DR-RAG: Addressing Retrieval Misalignment in Low-Resource Urdu Question Answering</i>	
Saad Ahmad, Muhammad Hammad, Muhammad Zeeshan, Faizad Ullah and Asim Karim	49
<i>Cross-Domain Evaluation of Transformer-Based Models for Punjabi Speech Emotion Recognition</i>	
Fatima Tu Zahra, Kulsoom Asim, Sandesh Kumar and Abdul Samad	59
<i>SiPaKosa: A Comprehensive Corpus of Canonical and Classical Buddhist Texts in Sinhala and Pali</i>	
Ranidu Hansaka Gurusinghe and Nevidu Jayatilleke	68
<i>BNLI: A Linguistically-Refined Bengali Dataset for Natural Language Inference</i>	
Farah Binta Haque, Md Yasin, Shishir Saha, Md Shoaib Akhter Rafi and Farig Sadeque	85
<i>Exploring Large Language Models for Multitask Learning in Bengali Text Classification</i>	
Md. Sajjad Hossain, Kawsar Ahmed, SUNY MD ASHRAF KHAN and Mohammed Moshuiul Hoque	92
<i>Nwāchā Munā: A Devanagari Speech Corpus and Proximal Transfer Benchmark for Nepal Bhasha ASR</i>	
Rishikesh Kumar Sharma, Safal Narshing Shrestha, Jenny Poudel, Rupak Tiwari, Arju Shrestha, Rupak Raj Ghimire and Bal Krishna Bal	105
<i>From Romanized to Devanagari: Enhancing Nepali Sentiment Analysis with NepaliXlit</i>	
Suraj Patel, Kashish Kumari Dhami, Norden Sherpa and Supriya Khadka	115
<i>Why Does Low-Rank Adaptation Work for Hindi-English Code-Mixing? A Geometric Analysis</i>	
Shashank Vishwakarma and Rakesh Kumar	127

<i>Hi-SEMFLOW: Lie Algebra–Based Semantic Flow for Span-Level Informal Language Identification in Hindi</i>	
Manikandan Ravikiran, Tanmay Tiwari, Vibhu Gupta and Rohit Saluja	137
<i>NeCCo: Nepali Cultural Commonsense Benchmark for Large Language Model Evaluation</i>	
Sanket Shrestha, Raunak Regmi, Sadikshya Ghimire, Satyam Rana and Supriya Khadka	154
<i>Reward-Guided Fine-Tuning of Whisper for Low-Resource Nepali Speech Recognition</i>	
Aadarsh Pandit, Yudhin Khanal, Ishan Pandey, Kushal Kunwar and Sunil Regmi	169
<i>Evaluating Linguistic Knowledge of LLMs in Tamil: The ILAKKANAM Benchmark</i>	
Jeyarajalingam Varsha, Menan Velayuthan, Sumirtha Karunakaran, Rasan Nivethiga and Kengatharaiyer Sarveswaran	176
<i>A Feature-Fusion Ensemble Approach for Tamil Hate Speech Detection</i>	
Sathasivam Nerujan and Kengatharaiyer Sarveswaran	190
<i>Comparative Analysis of Tokenizers in Tamil Text Classification in Low Resource Settings</i>	
Gokulan Sivakumaran, Randil Pushpananda and ERAD Bandara	198
<i>Improving Public Health Safety in Low-Resource Languages Using a Human-Verified Health Misinformation Corpus and Large Language Models</i>	
Sujal Maharjan, Astha Shrestha, Laxmi Thapa, Sweta Poudel, Shuvam Shiwakoti, Rabin Thapa, Kritesh Rauniyar and Surendrabikram Thapa	209
<i>Multimodal Hate and Sentiment Understanding in Low-Resource Text-Embedded Images for Online Safety and Digital Well-being</i>	
Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal and Usman Naseem	219
<i>Unigoa@CHiPSAL 2026: Early vs Late Fusion for Multimodal Hate and Sentiment Detection in Nepali Memes</i>	
Ashweta Fondekar, Milind Shivolkar and Jyoti Pawar	229
<i>HasNat@CHiPSAL 2026: Multimodal Hate Speech Detection in Low-Resource Nepali Memes Using Aligned Vision–Language Models</i>	
Alvee Hasan Chowdhury, MD. ABUL HASNAT and Adnan Faisal	237
<i>TeamHerald@CHiPSAL 2026: Hate Speech Detection and Sentiment Analysis of Nepali Memes Using Transformer-based Architectures and Ensemble Learning</i>	
Ashish Acharya, Anish Khatiwada, Rohit Khadka and Pragya Aryal	244
<i>NeuralNoodles@CHiPSAL 2026: Late-Fusion Multimodal Stacking for Nepali Meme Sentiment Classification</i>	
Sidratul Muntaha, Sabila Anzum, Arpita Mallik and Hasan Murad	250
<i>Multi-Modal-Minds@CHiPSAL 2026: A Comparative Study of Textual, Visual and Multimodal Architecture for Nepali Meme Moderation</i>	
Sandesh Shrestha, Bikram K.C., Akshyat Shah, Ashish Acharya and Rabin Thapa ...	258

<i>linus@CHiPSAL 2026: Multimodal Hate Speech and Sentiment Detection in Low-Resource Memes Using Late-Fusion Hybrid Architecture</i>	
Sunil Regmi, Bipesh Subedi, Saugat Singh and Suman Shrestha	267
<i>ZeroR@CHiPSAL 2026: Two-Stage Vision-Language Adaptation with Contrastive Learning for Nepali Meme Classification</i>	
Nitiz Khanal	275
<i>Team Oryu@CHiPSAL 2026: Integrating Text and Vision Transformers for Multimodal Hate Speech Detection in Memes</i>	
Noore Tamanna Orny, Joyeta Barua Moni, Md. Abtahee Kabir and Hasan Murad	284
<i>Cuet Yet Another Baseline@CHiPSAL LREC 2026: Shared Task on Multimodal Sentiment Understanding in Low-Resource Memes</i>	
Rotna Dipika Debnath, Shahrin Afroz Hoque Ruhi, Ayesha Labiba, Arpita Mallik and Hasan Murad	292
<i>MEME-Fusion@CHiPSAL 2026: Multimodal Ablation Study of Hate Detection and Sentiment Analysis on Nepali Memes</i>	
Samir Wagle, Reewaj Khanal and Abiral Adhikari	300
<i>eGrantha.ai@CHiPSAL 2026: Stochastic Image Captioning for Robust Hate Speech Detection in Low-Resource Nepali Memes</i>	
Anish Thapaliya	308
<i>EthosAI@CHiPSAL2026: Hate and Sentiment Understanding in Low-Resource Memes Using a Multimodal Approach</i>	
Vinayak Bansal, Deepawali Sharma, Aakash Singh and Vivek Kumar Singh	316

CHiPSAL2026 Workshop Program

16 May 2026

14:00 Workshop overview

Findings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026)

Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Ashwini Vaidya, Tafseer Ahmed and Bal Krishna Bal

14:15–14:40 Keynote address by Dr Usman Naseem, Macquarie University, Australia

14:40 Session 1

Chair: Professor Miriam Butt

Development of Burushaski Speech - English Text Translation Dataset

Tauqeer Saleem, Abdul Samad, Azkaa Nasir, Adina Adnan Mansoor, Fatima Faisal and Mahrukh Yousuf

Lost the Negation or Lost in Negation

Vennela Bairi and Parameswari Krishnamurthy

A Morphological Transducer for the Limbu Language

Avyaya Singh and Jonathan N. Washington

Evaluating Large Language Models for Medical Named Entity Recognition in Urdu: A Benchmark Study

Bushra Nasim, Kinza Latif, Muhammad Zohair, Muhammad Hassan Asif and Zarmeen Nasim

DR-RAG: Addressing Retrieval Misalignment in Low-Resource Urdu Question Answering

Saad Ahmad, Muhammad Hammad, Muhammad Zeeshan, Faizad Ullah and Asim Karim

Cross-Domain Evaluation of Transformer-Based Models for Punjabi Speech Emotion Recognition

Fatima Tu Zahra, Kulsoom Asim, Sandesh Kumar and Abdul Samad

SiPaKosa: A Comprehensive Corpus of Canonical and Classical Buddhist Texts in Sinhala and Pali

Ranidu Hansaka Gurusinghe and Nevidu Jayatilleke

BNLI: A Linguistically-Refined Bengali Dataset for Natural Language Inference

Farah Binta Haque, Md Yasin, Shishir Saha, Md Shoaib Akhter Rafi and Farig Sadeque

16 May 2026 (continued)

Exploring Large Language Models for Multitask Learning in Bengali Text Classification

Md. Sajjad Hossain, Kawsar Ahmed, SUNY MD ASHRAF KHAN and Mohammed Moshiul Hoque

16:00–16:30

Keynote address by Professor Monojit Choudhury - "The Curious Case of Honorifics in South Asian Languages and Their Treatment in LLMs"

16:30

Session 2

Chair: Professor Monojit Choudhury, Mohamed bin Zayed University of Artificial Intelligence

Nwāchā Munā: A Devanagari Speech Corpus and Proximal Transfer Benchmark for Nepal Bhasha ASR

Rishikesh Kumar Sharma, Safal Narshing Shrestha, Jenny Poudel, Rupak Tiwari, Arju Shrestha, Rupak Raj Ghimire and Bal Krishna Bal

From Romanized to Devanagari: Enhancing Nepali Sentiment Analysis with NepaliXlit

Suraj Patel, Kashish Kumari Dhami, Norden Sherpa and Supriya Khadka

Why Does Low-Rank Adaptation Work for Hindi-English Code-Mixing? A Geometric Analysis

Shashank Vishwakarma and Rakesh Kumar

Hi-SEMFLOW: Lie Algebra–Based Semantic Flow for Span-Level Informal Language Identification in Hindi

Manikandan Ravikiran, Tanmay Tiwari, Vibhu Gupta and Rohit Saluja

NeCCo: Nepali Cultural Commonsense Benchmark for Large Language Model Evaluation

Sanket Shrestha, Raunak Regmi, Sadikshya Ghimire, Satyam Rana and Supriya Khadka

Reward-Guided Fine-Tuning of Whisper for Low-Resource Nepali Speech Recognition

Aadarsh Pandit, Yudhin Khanal, Ishan Pandey, Kushal Kunwar and Sunil Regmi

16 May 2026 (continued)

17:15

Session 3

Chair: TBC

Evaluating Linguistic Knowledge of LLMs in Tamil: The ILAKKANAM Benchmark

Jeyarajalingam Varsha, Menan Velayuthan, Sumirtha Karunakaran, Rasan Nivethiga and Kengatharaiyer Sarveswaran

A Feature-Fusion Ensemble Approach for Tamil Hate Speech Detection
Sathasivam Nerujan and Kengatharaiyer Sarveswaran

Comparative Analysis of Tokenizers in Tamil Text Classification in Low Resource Settings

Gokulan Sivakumaran, Randil Pushpananda and ERAD Bandara

Improving Public Health Safety in Low-Resource Languages Using a Human-Verified Health Misinformation Corpus and Large Language Models

Sujal Maharjan, Astha Shrestha, Laxmi Thapa, Sweta Poudel, Shuvam Shiwakoti, Rabin Thapa, Kritesh Rauniyar and Surendrabikram Thapa

17:45

Findings - Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes@CHIPSAL2026

Multimodal Hate and Sentiment Understanding in Low-Resource Text-Embedded Images for Online Safety and Digital Well-being

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal and Usman Naseem

17:55

Concluding remarks

