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

Kengatharaiyer Sarveswaran[1], Surendrabikram Thapa[2], Ashwini Vaidya[3], Tafseer Ahmed[4], Bal Krishna Bal[5]

[1]University of Jaffna, Sri Lanka, [2]Virginia Tech, USA, [3]Indian Institute of Technology Delhi, India, [4]Mohammad Ali Jinnah University, Pakistan, [5]Kathmandu University, Nepal

Abstract

This paper presents the findings of the second workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026), held as part of LREC 2026. South Asia is one of the most linguistically diverse regions in the world, yet its languages remain severely underrepresented in language resources and technologies, particularly in the era of large language models (LLMs). The workshop brings together research addressing key challenges in this space, including data scarcity, morphological complexity, code-mixing, script diversity, and the lack of culturally grounded evaluation benchmarks. The workshop received 57 submissions, covering a wide range of languages, tasks, and modalities, including both widely spoken languages (e.g., Bengali, Hindi, Tamil, and Urdu) and extremely low-resource and endangered languages such as Burushaski, Limbu, and Nepal Bhasha (Newari). Several contributions introduce arguably first-of-their-kind resources and benchmarks for these languages, spanning both text and speech domains, and focusing on linguistically informed and culturally grounded data creation. In addition to the main track, the workshop hosted a shared task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes for Nepali, attracting strong community participation. The results highlight the effectiveness of multimodal approaches while also revealing persistent challenges in modelling culturally nuanced and low-resource data. Across the accepted papers and shared task, key insights include the central role of high-quality data, the limitations of current multilingual models in low-resource settings, and the need for culturally aware and data-centric approaches. Overall, CHiPSAL 2026 demonstrates the growing momentum in South Asian language processing and highlights the importance of sustained, community-driven efforts to build inclusive and representative language technologies.

Keywords: South Asian languages, low-resource NLP, linguistic resources, multimodal learning, shared task

1. Introduction

South Asia—comprising Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka—is home to nearly one-quarter of the world's population and represents one of the most linguistically diverse regions globally. The region includes languages from at least five major language families, along with several putative isolates, and comprises around 700 languages (Hock and Bashir, 2016). These languages are written in more than 25 scripts, reflecting a rich and complex linguistic ecology. In addition, a global diaspora exceeding 50 million speakers further extends the use of South Asian languages beyond the region. Despite this diversity and demographic significance, South Asian languages remain severely underrepresented in language resources and technologies.

Recent advances in large language models (LLMs) have led to significant progress in natural language processing; however, South Asian languages remain marginally represented in these models due to the limited availability of high-quality, curated data. While Unicode standardisation has largely addressed encoding issues, challenges persist in rendering, normalisation, and input methods, particularly for complex scripts and low-resource settings. Furthermore, the linguistic characteristics of South Asian languages—including rich morphology (Sarveswaran et al., 2021), diverse writing systems, code-mixing, and strong dialectal variation—pose substantial challenges for NLP. Closely related languages and shared scripts further complicate modelling and evaluation, often leading to performance disparities and unintended biases in multilingual systems.

The CHiPSAL (Challenges in Processing South Asian Languages) workshop addresses these challenges by bringing together research that focuses on linguistic, cultural, and technological aspects of processing South Asian languages. In particular, it emphasises the development of high-quality linguistic resources, culturally grounded evaluation benchmarks, and methods tailored to low-resource settings. By fostering collaboration across diverse language communities and research groups, the workshop aims to advance the state of NLP for South Asian languages while supporting the preservation and computational inclusion of their linguistic and cultural heritage. In this paper, we provide an overview of the accepted papers, shared tasks, and key directions emerging from the workshop.

This workshop represents the second edition of the CHiPSAL series, building on the success of the first workshop (Sarveswaran et al., 2025), which was co-located with COLING 2025 and brought to-

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 1–8
16 May 2026.

gether a growing community of researchers and practitioners working on South Asian language technologies.

2. Submission and Review processes

The workshop proposal was accepted by the LREC workshop chairs, after which the call for papers was disseminated through various channels, including mailing lists, direct contact emails, and the official workshop website[1]. In addition to the main workshop, a shared task was organised to further engage the research community and encourage broader participation and contributions towards South Asian language processing.

We received a total of 42 submissions to the main track of the CHiPSAL workshop. Of these, 2 papers were withdrawn by the authors, and the remaining submissions were considered for review. Following the review process, 21 papers were accepted for presentation, resulting in an acceptance rate of approximately 50%. Subsequently, 2 accepted papers were withdrawn, and 19 papers are included in the final proceedings.

All submissions underwent a rigorous peer-review process, with each paper evaluated by three program committee members to ensure a fair and thorough assessment. In total, 55 program committee members from academia and industry worldwide contributed to the review process. Among the accepted papers, selections for presentation were made to ensure coverage of diverse tasks and languages while accommodating the workshop schedule. The accepted papers were presented as oral presentations within the allocated time.

In addition to the main track, the workshop hosted a shared task on "Multimodal Hate and Sentiment Understanding in Low-Resource Memes," comprising two subtasks: (1) hate speech detection and (2) sentiment analysis. The shared task attracted participation from 23 teams, resulting in 14 system description paper submissions, of which 12 papers were accepted for publication.

3. Overview of Submissions

The workshop includes several contributions that extend computational research to extremely underrepresented and endangered languages by introducing foundational resources and tools. These include a structured Burushaski–English speech translation dataset comprising 10 hours of curated audio from 42 speakers, designed for an oral language with complex morphology (Saleem et al., 2026); the first morphological transducer for Limbu, enabling computational analysis for an endangered Sino-Tibetan language (Singh and Washington, 2026); and a 5.39-hour annotated speech corpus with an accompanying ASR benchmark for Nepal Bhasha (Newari), establishing initial infrastructure for speech technology in this language (Sharma et al., 2026). These works collectively contribute new linguistic coverage and resource development in areas that have remained largely absent from prior NLP research.

The papers present a substantial body of work on the creation of high-quality, task-specific datasets and benchmarks that are explicitly grounded in linguistic structure and cultural context, moving beyond opportunistic or purely scraped corpora. BNLI introduces a linguistically curated Bengali NLI dataset designed to address annotation errors, ambiguity, and lack of diversity through a controlled annotation pipeline (Haque et al., 2026)[2] . Similarly, NEGTEG provides a structured benchmark for Telugu comprising five tasks—negation detection, translation, paraphrase detection, sentiment analysis, and polarity flipping—enabling systematic evaluation of a core linguistic phenomenon (Bairi and Krishnamurthy, 2026)[3] . NeCCo extends this direction by introducing a culturally grounded Nepali commonsense benchmark spanning five domains, including kinship, rituals, and idiomatic expressions, thereby incorporating culturally embedded knowledge into evaluation (Shrestha et al., 2026b)[4] . In the Tamil context, ILAKKANAM constructs a benchmark of 820 linguistically annotated school-level questions across multiple linguistic categories, enabling fine-grained evaluation of linguistic competence (Varsha et al., 2026). Complementing these, Nep-Health-Misinfo provides a human-verified corpus for health misinformation detection in Nepali, constructed through a machine translation post-editing pipeline involving native experts (Maharjan et al., 2026)[5].

In addition to task-specific benchmarks, the workshop also introduces large-scale and diverse corpora that significantly expand data availability for low-resource languages across modalities and domains. SiPaKosa contributes a substantial Sinhala–Pali corpus comprising 786K sentences and 9.25M words drawn from 16 historical Buddhist documents and canonical sources, combining OCR-processed manuscripts with curated web data (Gursinghe and Jayatilleke, 2026)[6]. For Nepali, a Romanized social media dataset of over 6,500 comments, including a balanced subset of 1,518 manually annotated instances, supports the study

[1] https://sites.google.com/view/chipsal/

[2] BNLI-Dataset

[3] NEGTEG Dataset

[4] NeCCo Dataset

[5] Nep-Health-Misinfo Dataset

[6] SiPaKosa Dataset

Language	Language Family	Papers
Bengali	Indo-Aryan	(Haque et al., 2026; Hossain et al., 2026)
Burushaski	Language Isolate	(Saleem et al., 2026)
Hindi	Indo-Aryan	(Ravikiran et al., 2026; Vishwakarma and Kumar, 2026)
Limbu	Sino-Tibetan	(Singh and Washington, 2026)
Nepal Bhasha (Newari)	Sino-Tibetan	(Sharma et al., 2026)
Nepali	Indo-Aryan	(Shrestha et al., 2026b; Patel et al., 2026; Maharjan et al., 2026; Pandit et al., 2026)
Pali	Indo-Aryan	(Gururisinghe and Jayatilleke, 2026)
Punjabi	Indo-Aryan	(Zahra et al., 2026)
Sinhala	Indo-Aryan	(Gururisinghe and Jayatilleke, 2026)
Tamil	Dravidian	(Nerujan and Sarveswaran, 2026; Sivakumaran et al., 2026; Varsha et al., 2026)
Telugu	Dravidian	(Bairi and Krishnamurthy, 2026)
Urdu	Indo-Aryan	(Nasim et al., 2026; Ahmad et al., 2026)

Table 1: Languages covered in CHiPSAL 2026 regular (non–shared task) papers.

of cross-script and informal language processing (Patel et al., 2026)[7]. Speech resources are also represented: a Punjabi speech emotion dataset covering four emotion classes is introduced alongside a multi-strategy evaluation framework (Zahra et al., 2026)[8], while additional datasets include a publicly released Nepali ASR corpus used for reward-guided training (Pandit et al., 2026)[9] and an Urdu news headline dataset annotated for medical named entity recognition (Nasim et al., 2026)[10] .

4. Summaries of Accepted Papers

The BNLI dataset addresses limitations in Bengali NLI resources such as annotation errors, ambiguity, and lack of diversity. It is constructed with a rigorous annotation pipeline ensuring semantic clarity and balanced entailment, contradiction, and neutrality classes. Benchmarking with multilingual and Bengali-specific transformer models demonstrates improved reliability and interpretability, establishing BNLI as a strong foundation for Bengali and other low-resource inference tasks (Haque et al., 2026).

For Tamil hate speech detection, a hybrid framework combines L3Cube-TamilBERT representations with FastText embeddings to address subword fragmentation. Using Last-4 layer averaging and dual pooling (mean + max), the model achieves a Macro-F1 of 0.7883 and Hate Recall of 0.7503. A stacking ensemble further achieves Hate Precision of 0.9296, demonstrating improved detection performance (Nerujan and Sarveswaran, 2026).

The Burushaski–English speech translation dataset introduces an audio-first methodology tailored to a morphologically complex, ergative-absolutive language with a four-gender agreement system. Using structured elicitation, crowdsourcing, and translation alignment, the dataset contains approximately 10 hours of curated audio from 42 speakers. Preliminary Whisper-based translation experiments are presented, but the primary contribution is a scalable framework for speech-first corpus design (Saleem et al., 2026).

Hi-SEMFLOW proposes a Lie algebra–based semantic flow framework for span-level labeling in Hindi. It models label refinement as a continuous process using antisymmetric generators. On the HiSlang-4.9k benchmark, it improves span-level F1 by 2–3 absolute points and yields consistent macro-F1 gains, demonstrating effectiveness over discrete structured decoding approaches (Ravikiran et al., 2026).

For Punjabi speech emotion recognition, a dataset covering four emotions (angry, happy, sad, neutral) is introduced. Three models are evaluated: CNN-BiLSTM, ResNet-34, and Wav2Vec 2.0. ResNet-34 achieves the best performance with 96% accuracy in the combined-domain setting (E4), while cross-domain evaluations highlight challenges, especially for neutral emotion classification (Zahra et al., 2026).

The NEGTEG benchmark for Telugu includes five tasks: negation detection, translation, paraphrase detection, sentiment analysis, and polarity flipping. Evaluation shows that multilingual models struggle significantly with Telugu negation across all tasks, particularly due to morphological complexity and linguistic variation (Bairi and Krishnamurthy, 2026).

NeCCo introduces a culturally grounded Nepali commonsense benchmark covering five domains: kinship and social hierarchy; festivals, rituals, and geography; idioms, proverbs, and metaphors; commonsense and daily life; and gastronomy, agriculture, and nature. Evaluation shows models per-

[7] Nepali Romanized social media dataset
[8] Kaggle Dataset
[9] Nepali ASR Dataset
[10] Urdu NewsHeadline Dataset

form better on globally documented knowledge but struggle with culturally dense tasks, exhibiting brittleness, hallucination, and difficulty with implicit reasoning (Shrestha et al., 2026b).

NepaliXlit improves transliteration from Romanized Nepali to Devanagari using 2,943 training word pairs and 736 test pairs. It improves transliteration accuracy by 8% and reduces character error rate by 11%. A dataset of 6,500 comments (with 1,518 annotated instances) shows that transliteration improves sentiment classification, while LLMs outperform encoder models in cross-script scenarios (Patel et al., 2026).

A Tamil tokenization study evaluates WordPiece, SentencePiece, and BBPE using a KaggleTamil news dataset. WordPiece and SentencePiece outperform BBPE in efficiency and classification performance. While BBPE eliminates OOV words, excessive fragmentation harms learning. Increasing vocabulary improves WordPiece and SentencePiece but not BBPE (Sivakumaran et al., 2026).

ILAKKANAM introduces a Tamil linguistic benchmark of 820 questions from Grades 1–13, annotated into five linguistic and factual categories. Evaluation shows Gemini 2.5 performs best overall, while open-source models lag behind. Performance declines with increasing grade level and linguistic complexity, and no strong correlation exists between overall performance and linguistic category identification (Varsha et al., 2026).

Nep-Health-Misinfo is introduced as the first human-verified Nepali health misinformation corpus, built from four existing benchmarks using an MTPE pipeline with native experts. The study finds a strong translation asymmetry, where SOTA models performs much better on factual health content than on deceptive narratives.Translation evaluation shows BLEU drops from 43.21 (factual) to 19.11 (deceptive), with TER reaching 62.42. Few-shot prompting improves Macro-F1 from 0.7188 (zero-shot) to 0.8488 for Qwen2.5-7B, though results remain sensitive to exemplar selection (Maharjan et al., 2026).

A Bengali multitask classification study evaluates LLMs across five domains. Results show chain-of-thought prompting often degrades performance. Gemma-3-4B achieves the most balanced performance across both in- context learning and parameter-efficient fine-tuning settings (Hossain et al., 2026).

For Urdu medical NER, ChatGPT-4o and LLAMA 3.2 are evaluated on 2,057 annotated headlines across five entity types. ChatGPT-4o achieves F1 0.35 (disease), LLAMA 3.2 achieves 0.33, while treatment F1 scores are extremely low (0.011 and 0.036). Overall micro-F1 scores are 0.187 and 0.183, indicating poor performance (Nasim et al., 2026).

The Nwāchā Munā corpus introduces 5.39 hours of Nepal Bhasha speech data. Fine-tuning a Nepali Conformer reduces CER from 52.54% to 17.59%, matching Whisper-Small performance with fewer parameters, demonstrating effectiveness of proximal cross-lingual transfer (Sharma et al., 2026).

The first morphological transducer for Limbu achieves 60% coverage, 88% precision, and 32% recall using a small lexicon derived from field data and Bible translations. It highlights the need for expanded lexical resources and community involvement (Singh and Washington, 2026).

SiPaKosa introduces a corpus of 786K sentences and 9.25M words from 16 historical documents and canonical texts. Evaluation across 10 models shows perplexity ranges from 1.09 to 189.67, with proprietary models outperforming open-source ones by 3–6 times (Gurursinghe and Jayatilleke, 2026).

A reward-guided fine-tuning approach for Nepali ASR uses 2,000 human-rated samples to train a classifier with 81% accuracy to filter bad samples. Filtering twice and retraining ASR on 40,000 clip training subset drawn from a 68.4 hour corpus improves WER from 5.60% to 4.89% and CER from 5.10% to 4.52%, yielding 11–13% relative gains over baseline (Pandit et al., 2026).

A study on LoRA for Hindi-English code-mixing through spectral analysis of mBERT and MuRIL shows pre-trained attention ranks of 437–441, while LoRA updates have ranks 2.1–5.9, achieving $136\times$ compression. CKA scores show alignment of 0.279 (Hindi) vs. 0.093 (English), with statistical significance (Wilcoxon $p < 10^{-19}$), supporting low-dimensional subspace hypotheses (Vishwakarma and Kumar, 2026).

DR-RAG introduces dual-representation retrieval (text chunks + QA pairs) for Urdu. It improves Urdu METEOR by 38×, ROUGE-1 by 140%, and reduces latency by 43%. LLM-as-judge scores improve from 1.93 to 3.03 (faithfulness) and 2.21 to 2.99 (overall quality), demonstrating effectiveness of representation alignment between queries and indexed content.(Ahmad et al., 2026).

5. Key Takeaways

1. **Severe Resource Scarcity Remains the Central Bottleneck.** Many works highlight the lack of high-quality datasets across South Asian languages such as Bengali, Tamil, Nepali, Urdu, Punjabi, and Burushaski, which continues to hinder model development and evaluation (Haque et al., 2026; Nerujan and Sarveswaran, 2026; Saleem et al., 2026; Zahra et al., 2026; Nasim et al., 2026; Shrestha et al., 2026b).

2. **Dataset Quality and Linguistic Curation Are**

Critical. Carefully designed and linguistically grounded datasets significantly improve reliability, interpretability, and evaluation of models (Haque et al., 2026; Shrestha et al., 2026b; Varsha et al., 2026; Maharjan et al., 2026; Bairi and Krishnamurthy, 2026).

3. **Multimodal Understanding is Essential but Challenging.** Meme-based communication requires integrating textual and visual signals, and unimodal approaches fail to capture full semantics in low-resource settings (Bansal et al., 2026; Orny et al., 2026; Regmi et al., 2026; Shrestha et al., 2026a).

4. **Fusion Strategies Strongly Influence Performance.** The effectiveness of early fusion, late fusion, cross-modal attention, stacking, and voting strategies varies depending on task type and modality interaction (Acharya et al., 2026; Fondekar et al., 2026; Wagle et al., 2026; Muntaha et al., 2026).

5. **Low-Resource Settings Expose Weaknesses of Current Models.** State-of-the-art models struggle with linguistic phenomena such as negation, cultural reasoning, and domain-specific tasks, indicating shallow understanding in low-resource contexts (Bairi and Krishnamurthy, 2026; Shrestha et al., 2026b; Nasim et al., 2026; Hossain et al., 2026).

6. **Morphological Complexity and Tokenization Are Major Issues.** Rich morphology and subword fragmentation significantly affect representation learning and downstream performance (Nerujan and Sarveswaran, 2026; Sivakumaran et al., 2026; Ahmad et al., 2026).

7. **Code-Mixing and Script Variation Add Significant Complexity.** Mixed scripts and informal language usage create challenges for multilingual models, though transliteration and cross-lingual strategies provide improvements (Patel et al., 2026; Fondekar et al., 2026; Vishwakarma and Kumar, 2026).

8. **Speech and Oral Languages Require Different Paradigms.** For languages with limited written resources, speech-first approaches and audio-centric datasets are more suitable than text-based pipelines (Saleem et al., 2026; Sharma et al., 2026).

9. **Data-Centric and Lightweight Methods Are Effective Alternatives.** Improvements can be achieved through data filtering, human-in-the-loop approaches, proximal transfer, and parameter-efficient methods (Pandit et al., 2026; Sharma et al., 2026; Hossain et al., 2026).

10. **Cultural and Contextual Knowledge Is Crucial for True Language Understanding.** Models trained on global datasets fail to capture culturally embedded knowledge, leading to poor performance in culturally grounded tasks (Shrestha et al., 2026b).

6. Shared Task Description

In addition to the main track, the workshop hosted a shared task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes, focusing on monolingual Nepali memes written in the Devanagari script. The task comprised two subtasks: (1) hate speech detection and (2) sentiment analysis, both requiring joint reasoning over textual and visual modalities. The shared task attracted strong participation from the community, with 65 registered participants leading to 23 teams submitting systems for hate detection and 43 registered participants with 13 teams submitting systems for sentiment analysis. The diversity of submissions reflects a wide range of approaches, including multimodal fusion architectures, caption-based methods using vision-language models, and ensemble techniques. Further details on the dataset, task design, and evaluation protocol are provided in (Thapa et al., 2026).

The submitted systems demonstrated that multimodal approaches are generally more effective than unimodal baselines in low-resource settings. The top-performing system achieved macro-F1 scores of 80.52% for hate speech detection and 68.81% for sentiment analysis using a late-fusion hybrid architecture combining strong multilingual text encoders with vision models and employing discriminative learning rates. Overall, the results indicate that hate detection is relatively more tractable than sentiment analysis, which requires capturing finer-grained semantic and cultural nuances. Despite promising performance, the task also highlights persistent challenges, including the limited availability of Devanagari-centric pretrained models and the difficulty of modelling culturally grounded multimodal content, pointing to important directions for future research.

7. Conclusion and Future Directions

The second edition of the CHiPSAL workshop demonstrates the growing interest and momentum in research on South Asian language processing. The workshop received a 57 of submissions, covering a wide range of languages, tasks, and modalities. Notably, the accepted papers span both widely spoken and underrepresented languages, including Bengali, Hindi, Nepali, Pali, Punjabi, Sinhala,

Tamil, Telugu, and Urdu, as well as extremely low-resource and endangered languages such as Burushaski, Limbu, and Nepal Bhasha (Newari). Several contributions focused on resource-intensive efforts, as outlined in Section 3, including the development of datasets and tools for these languages across both text and speech domains.

These contributions represent significant progress when considered in the context of limited funding, infrastructural constraints, and organisational challenges that characterise research in this region. These efforts are particularly valuable, as they lay the groundwork for future research and resource development, contributing to a more inclusive and representative NLP landscape.

The strong response to the workshop, with 57 submissions for a half-day event, highlights the importance and timeliness of initiatives such as CHiPSAL. At the same time, it is important to acknowledge that participation in major international venues remains challenging for many researchers in the region due to financial, institutional, and organisational constraints. Addressing these barriers is crucial for ensuring broader and more equitable participation in the global research community.

In light of this, we plan to expand CHiPSAL through regional initiatives to better support local research communities. The first such event is planned to be held in Sri Lanka on 31 July 2026. We invite researchers and institutions across South Asia to collaborate in organising similar satellite events, with the aim of strengthening regional networks and fostering sustained and inclusive research engagement in South Asian language technologies.

8. Acknowledgments

We would like to thank Ahrane Mahaganapathy, Ajintha Sivakulasingam, and Nishanthini Kanthakumar for their valuable assistance as volunteers in organising the workshop.

We are grateful to all members of the programme committee for their timely reviews despite the short timeframe; their contributions have significantly improved the quality of the workshop.

We also thank the LREC workshop chairs for accepting our proposal and for their support in facilitating the successful organisation of the workshop.

9. Bibliographical References

Ashish Acharya, Anish Khatiwada, Rohit Khadka, and Pragya Aryal. 2026. TeamHerald@CHIPSAL 2026: Hate Speech Detection and Sentiment Analysis of Nepali Memes using Transformer-based Architectures and Ensemble Learning. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Saad Ahmad, Muhammad Hammad, Muhammad Zeeshan, Faizad Ullah, and Asim Karim. 2026. DR-RAG: Addressing Retrieval Misalignment in Low-Resource Urdu Question Answering. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Vennela Bairi and Parameswari Krishnamurthy. 2026. Lost the Negation or Lost in Negation. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Vinayak Bansal, Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2026. EthosAI@CHiPSAL2026: Hate and Sentiment Understanding in Low-Resource Memes using a Multimodal Approach. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Ashweta A. Fondekar, Milind M. Shivolkar, and Jyoti D. Pawar. 2026. Unigoa@CHiPSAL 2026: Early vs Late Fusion for Multimodal Hate and Sentiment Detection in Nepali Memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Ranidu Gurursinghe and Nevidu Jayatilleke. 2026. SiPaKosa: A Comprehensive Corpus of Canonical and Classical Buddhist Texts in Sinhala and Pali. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Farah Binta Haque, Md Yasin, Shishir Saha, Md Shoaib Akhter Rafi, and Farig Sadeque. 2026. BNLI: A Linguistically-Refined Bengali Dataset for Natural Language Inference. In *Proceedings of the Second Workshop on Challenges*

in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Hans Henrich Hock and Elena Bashir, editors. 2016. *The Languages and Linguistics of South Asia*. De Gruyter Mouton, Berlin, Boston.

Md. Sajjad Hossain, Kawsar Ahmed, Suny Md Ashraf Khan, and Mohammed Moshiul Hoque. 2026. Exploring Large Language Models for Multitask Learning in Bengali Text Classification. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Sujal Maharjan, Astha Shrestha, Laxmi Thapa, Sweta Poudel, Shuvam Shiwakoti, Rabin Thapa, Kritesh Rauniyar, and Surendrabikram Thapa. 2026. Improving Public Health Safety in Low-Resource Languages Using a Human-Verified Health Misinformation Corpus and Large Language Models. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Sidratul Muntaha, Sabila Anzum, Arpita Mallik, and Hasan Murad. 2026. NeuralNoodles@CHiPSAL 2026: Late-Fusion Multimodal Stacking for Nepali Meme Sentiment Classification. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Bushra Nasim, Kinza Latif, Muhammad Zohair, Muhammad Hassan Asif, and Zarmeen Nasim. 2026. Evaluating Large Language Models for Medical Named Entity Recognition in Urdu: A Benchmark Study. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Sathasivam Nerujan and Kengatharaiyer Sarveswaran. 2026. A Feature-Fusion Ensemble Approach for Tamil Hate Speech Detection. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Noore Tamanna Orny, Joyeta Barua Moni, Md. Abtahee Kabir, and Hasan Murad. 2026. Team Oryu@CHiPSAL 2026: Integrating Text and Vision Transformers for Multimodal Hate Speech Detection in Memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Aadarsh Pandit, Yudhin Khanal, Ishan Pandey, Kushal Kunwar, and Sunil Regmi. 2026. Reward-Guided Fine-Tuning of Whisper for Low-Resource Nepali Speech Recognition. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Suraj Patel, Kashish Kumari Dhami, Norden Sherpa, and Supriya Khadka. 2026. From Romanized to Devanagari: Enhancing Nepali Sentiment Analysis with NepaliXlit. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Manikandan Ravikiran, Tanmay Tiwari, Vibhu Gupta, and Rohit Saluja. 2026. Hi-SEMFLOW: Lie Algebra–Based Semantic Flow for Span-Level Informal Language Identification in Hindi. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Sunil Regmi, Bipesh Subedi, Saugat Singh, and Suman Shrestha. 2026. linus@CHiPSAL 2026: Multimodal Hate Speech and Sentiment Detection in Low-Resource Memes using Late-Fusion Hybrid Architecture. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Tauqeer Saleem, Abdul Samad, Azkaa Nisar, Fatima Faisal, Adina Adnan Mansoor, and Mahrukh Yousuf. 2026. Development of Burushaski Speech – English Text Translation Dataset. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2021. Thamizhi morph: A morpho-

logical parser for the tamil language. *Machine Translation*, 35(1):37–70.

Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Sana Shams, Ashwini Vaidya, and Bal Krishna Bal. 2025. A brief overview of the first workshop on challenges in processing South Asian languages (CHiPSAL). In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 1–8, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Rishikesh Kumar Sharma, Safal Narshing Shrestha, Jenny Poudel, Rupak Tiwari, Arju Shrestha, Rupak Raj Ghimire, and Bal Krishna Bal. 2026. Nwāchā Munā: A Devanagari Speech Corpus and Proximal Transfer Benchmark for Nepal Bhasha ASR. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Sandesh Shrestha, Bikram K.C., Akshyat Shah, Ashish Acharya, and Rabin Thapa. 2026a. Multi-Modal-Minds@CHiPSAL 2026: A Comparative Study of Textual, Visual and Multimodal Architecture for Nepali Meme Moderation. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Sanket Shrestha, Raunak Regmi, Sadikshya Ghimire, Satyam Rana, and Supriya Khadka. 2026b. NeCCo: Nepali Cultural Commonsense Benchmark for Large Language Model Evaluation. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Avyaya Singh and Jonathan North Washington. 2026. A Morphological Transducer for the Limbu Language. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Gokulan Sivakumaran, Randil Pushpananda, and ERAD Bandara. 2026. Comparative Analysis of Tokenizers in Tamil Text Classification in Low Resource Settings. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal, and Usman Naseem. 2026. Multimodal hate and sentiment understanding in low-resource text-embedded images for online safety and digital well-being. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Jeyarajalingam Varsha, Menan Velayuthan, Sumirtha Karunakaran, Rasan Nivethiga, and Kengatharaiyer Sarveswaran. 2026. Evaluating Linguistic Knowledge of LLMs in Tamil: The *ILAKKANAM* Benchmark. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Shashank Vishwakarma and Rakesh Kumar. 2026. Why Does Low-Rank Adaptation Work for Hindi-English Code-Mixing? A Geometric Analysis. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Samir Wagle, Reewaj Khanal, and Abiral Adhikari. 2026. MEME-Fusion@CHiPSAL 2026: Multimodal Ablation Study of Hate Detection and Sentiment Analysis on Nepali Memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).

Fatima Tu Zahra, Kulsoom Asim, Sandesh Kumar, and Abdul Samad. 2026. Cross-Domain Evaluation of Transformer-Based Models for Punjabi Speech Emotion Recognition. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) @ LREC 2026*, Palma, Mallorca (Spain). European Language Resources Association (ELRA).