A Brief Overview of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)

Kengatharaiyer Sarveswaran
University of Jaffna
Sri Lanka
sarves@univ.jfn.ac.lk

Surendrabikram Thapa
Virginia Tech, Blacksburg
United States of America
surendrabikram@vt.edu

Sana Shams
Al-Khawarizmi Institute
of Computer Science
UET, Pakistan
sana.shams@kics.edu.pk

Ashwini Vaidya
Indian Institute of Technology, Delhi
India
avaidya@iitd.ac.in

Bal Krishna Bal
Kathmandu University
Nepal
bal@ku.edu.np

Abstract

In this paper, we provide a brief summary of the inaugural workshop on Challenges in Processing South Asian Languages (CHiPSAL) held as part of COLING 2025. The workshop included regular papers, invited keynotes, and shared task papers, fostering a collaborative platform for exploring challenges in processing South Asian languages. The shared task focused on Devanagari-script language understanding, encompassing subtasks on language identification, hate speech detection, and target classification. This workshop series aims to address linguistic and cultural nuances, resource constraints, and orthographic complexities in low-resource South Asian languages while advancing NLP research and promoting multilingual inclusivity.

1 Introduction

South Asia, encompassing Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka, is one of the world's most populous regions, accounting for nearly a quarter of the global population (see Table 1). The region is linguistically diverse, featuring languages from at least four major language families and several potential linguistic isolates. (Hock and Bashir, 2016; Arora et al., 2022) With over 700 languages and approximately 25 major scripts, South Asia boasts a rich cultural and linguistic heritage. Furthermore, over 50 million South Asians reside abroad. Despite this remarkable diversity, South Asian languages remain significantly underrepresented in language technology.

Recent large language models (LLMs) incorporate limited data from South Asia, and the processing of South Asian languages presents significant challenges, starting with encoding issues. Although most scripts are standardized in Unicode, many applications fail to render them accurately due to orthographic complexities. Additionally, input methods for these languages remain a persistent barrier in the region. The linguistic complexity of South Asian languages, characterized by diverse writing systems and extensive literary traditions, further complicates natural language processing (NLP) tasks. Dialectal and cultural variations, along with the close linguistic relationships among these languages, introduce additional layers of complexity.

This workshop addresses the multifaceted challenges in processing South Asian languages, focusing on linguistic and cultural factors, encoding and orthographic issues, and resource constraints. By tackling these issues, we aim to advance NLP for South Asian languages while preserving and promoting their rich linguistic and cultural heritage. In this paper, we provide a brief overview of our accepted papers, shared tasks, and the workshop's future directions.

2 Submission and Review

We received 46 long papers, 13 short papers for the main track of CHiPSAL workshop, and 20 shared-task papers which is organised as the part of the workshop. Twelve papers were later withdrawn by authors, and two papers were desk-rejected.

We accepted a total of 38 papers for the work-

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 1–8
January 19, 2025. ©2025 International Committee on Computational Linguistics

Country	Population (millions)	Living Languages	Literacy Rate (%)
Afghanistan	38.347	33	43
Bangladesh	166.303	36	74
Bhutan	0.772	21	67
India	1380	424	74
Maldives	0.541	1	98
Nepal	30.226	109	68
Pakistan	225.2	68	59
Sri Lanka	22.156	5	92
Total	**1863.549**	**697**	-
	(23.43% of World)	(10.07% of World)	

Table 1: South Asian Languages and Literacy Data (Eberhard et al., 2024)

shop, including 3 short papers, 17 long papers, and 18 shared task papers. For the main workshop track, 48 submissions were considered for review, of which 20 were accepted, resulting in an acceptance rate of approximately 41.7%. Of these, 8 papers were selected for oral presentations, while the remaining 12 were designated for poster presentations. The selection for oral and poster presentations aimed to ensure coverage of diverse tasks and languages while accommodating all presentations within a single day. Each submission underwent a rigorous review process, with three program committee members evaluating each paper to ensure a fair and thorough assessment. Overall, 55 program committee members from around the world, representing both academia and industry, contributed to the review process.

The submissions cover research on Bengali, English, Hindi, Kannada, Malayalam, Nepali, Pashto, Punjabi, Sinhala, Tamil, Telugu and Urdu languages on topics e.g. low-resource language challenges, script and linguistic complexity, speech processing and recognition, hate speech and code-mixing, and linguistic resource development.

3 Accepted papers

3.1 Long papers

Thapa et al. (2025a), in *Development of Pre-Trained Transformer-based Models for the Nepali Language*, highlight the under-representation of Nepali in NLP due to limited resources and monolingual corpora. They address this by collecting 27.5 GB of Nepali text data and pre-training BERT, RoBERTa, and GPT-2 models. They also explore instruction tuning, improving performance on Nepali datasets. Their models surpass the NepgLUE benchmark by 2 points, scoring 95.60, and perform better on text generation tasks, advancing Nepali text processing.

Chavinda and Thayasivam (2025), in *A Dual Contrastive Learning Framework for Enhanced Hate Speech Detection in Low-Resource Languages*, address hate speech detection in Sinhala and Tamil. They introduce a framework combining Multilingual Large Language Models (MLLMs) with Dual Contrastive Learning (DCL) to enhance detection. Using datasets from Facebook and Twitter, their approach outperforms traditional models, with the Twitter/twhin-bert-base model showing the best results. This study advances hate speech detection in low-resource languages.

Dhakal and Baral (2025), in *Abstractive Summarization of Low-Resourced Nepali Language Using Multilingual Transformers*, apply mBART and mT5 models to summarize Nepali news headlines. They create a headline corpus and fine-tune the models with Low-Rank Adaptation (LoRA) and quantization. Evaluations show the 4-bit quantized mBART model performs best. This work advances abstractive summarization and NLP for Nepali.

Neupane et al. (2025), in *Structured Information Extraction from Nepali Scanned Documents Using Layout Transformer and LLMs*, develop methods for extracting information from Nepali documents. They use the Language Independent Layout Transformer (LiLT), achieving an F1 score of 0.87, and compare it with LLMs like GPT-4o and Llama 3.1 8B. Their findings provide a foundation for digitizing Nepali texts.

Duwal et al. (2025), in *Domain-Adaptive Continual Learning for Low-Resource Tasks: Evaluation on Nepali*, explore domain-adaptive pre-training

(DAPT) to improve Llama 3 8B for Nepali. Using synthetic data and 4-bit QLoRA, they evaluate performance and knowledge retention, finding a 19.29% improvement in higher-shot evaluations. This study highlights DAPT's potential for low-resource tasks.

Rahothvarman et al. (2025), in *Bridge the GAP: Multi-Lingual Models for Ambiguous Pronominal Coreference Resolution in South Asian Languages*, address coreference resolution in Dravidian languages by creating the mGAP dataset. They develop joint embedding and cross-attention models, demonstrating the latter's effectiveness in capturing pronoun-candidate relations and leveraging transfer learning for low-resource languages.

Dasari et al. (2025), in *Sandhi Splitting in Tamil and Telugu: A Sequence-to-Sequence Approach Leveraging Transformer Models*, tackle sandhi splitting by creating annotated corpora and implementing sequence-to-sequence transformers. Their models, evaluated on the IN22-Conv Benchmark, improve preprocessing for morphologically rich languages like Tamil and Telugu.

James and Krishnamurthy (2025), in *POS-Aware Neural Approaches for Word Alignment in Dravidian Languages*, explore neural methods like SimAlign and AWESOME-align for Tamil and Telugu. They show that POS-tag fine-tuning improves alignment accuracy by 6–7% and investigate cross-linguistic mappings with English, highlighting the complexities of low-resource language alignment.

Pokharel and Agrawal (2025), in *neDIOM: Dataset and Analysis of Nepali Idioms*, introduce a Nepali idioms dataset and evaluate multilingual models on processing figurative language. They find that smaller models outperform larger ones, offering a resource for advancing idiom processing in low-resource languages.

Tahir et al. (2025), in *Benchmarking the Performance of Pre-Trained LLMs Across Urdu NLP Tasks*, benchmark seven pre-trained LLMs across 17 Urdu NLP tasks. They find models with richer language-specific data, like Llama 3.1-8B, often outperform larger models in tasks, emphasizing the importance of linguistic diversity in NLP research.

Khalid et al. (2025), in *Bridging the Bandwidth Gap: A Mixed Band Telephonic Urdu ASR Approach with Domain Adaptation for Banking Applications*, presents a telephonic Urdu ASR system using a corpus of 445 speakers. Comparing GMM-HMM and TDNN models, they find TDNN outperforms GMM-HMM. Mixing narrow-band and wide-band speech reduces Word Error Rates (WER), and domain adaptation with a specialized lexicon enhances performance for banking applications.

Chhetri and Poudyal (2025), in *Impacts of Vocoder Selection on Tacotron-Based Nepali Text-To-Speech Synthesis*, evaluate WaveNet and MelGAN vocoders for Nepali TTS. The study uses Nepali OpenSLR and News male voice datasets. They find that Tacotron2 + MelGAN consistently outperforms Tacotron2 + WaveNet in naturalness and higher Mean Opinion Score (MOS).

Thevakumar et al. (2025), in *EmoTa: A Tamil Emotional Speech Dataset*, introduce a dataset of 936 utterances from 22 Tamil speakers expressing five emotions. Fleiss' Kappa shows substantial agreement (0.74), and machine learning models achieve F1-scores above 0.90 for emotion classification. EmoTa supports Tamil speech emotion recognition research.

Ghimire et al. (2025), in *Improving Accuracy of Low-Resource ASR Using Rule-Based Character Constituency Loss (RBCCL)*, introduce RBCCL to improve transcription in Devanagari script. Combining RBCCL with cross-entropy loss reduces Word Error Rate (WER) from 47.1% to 23.41%, enhancing low-resource ASR performance.

Thayasivam et al. (2025), in *SiTa - Sinhala and Tamil Speaker Diarization Dataset in the Wild*, introduce a dataset addressing the lack of conversational data for speaker diarization in Sinhala and Tamil. They benchmark existing models, providing a resource for advancing speaker diarization in low-resource languages.

Srivastava (2025), in *DweshVaani: An LLM for Detecting Religious Hate Speech in Code-Mixed Hindi-English*, proposes Dwesh-Vaani, a fine-tuned Gemma-2 model outperforming other approaches for detecting hate speech and religion-specific targets in code-mixed Hindi-English. The study highlights challenges and opportunities in this domain.

Tanjila et al. (2025), in *Bengali ChartSumm: A Benchmark Dataset and Study on Feasibility of Large Language Models on Bengali Chart-to-Text Summarization*, introduce a dataset with 4,100 Bengali charts and summaries. Evaluating models like mT5 and BanglaT5, they establish baselines to support low-resource NLP research.

3.2 Short papers

Karim and Uzuner (2025), in *Leveraging Machine-Generated Data for Joint Intent Detection and Slot Filling in Bangla: A Resource-Efficient Approach*, generated a Bangla dataset for Natural Language Understanding (NLU) by translating the English SNIPS dataset (Coucke et al., 2018) using the LLaMA-3 model, focusing on intent detection and slot-filling tasks. They evaluated both separate and joint modeling approaches using different BERT variants, finding that the multilingual BERT (mBERT) achieved the best performance, with 97.83% intent accuracy and 91.03% slot-filling F1 score.

Sehar et al. (2025), in *Benchmarking Whisper for Low-Resource Speech Recognition: An N-Shot Evaluation on Pashto, Punjabi, and Urdu*, benchmarked the Whisper ASR model's performance on three low-resource languages - Pashto, Punjabi, and Urdu - by first evaluating its zero-shot performance and then fine-tuning the Whisper Small model on domain-specific datasets. They found that few-shot fine-tuning significantly reduced the Word Error Rate (WER), with improvements ranging from 6-19 percentage points across different languages and datasets, demonstrating the potential of adapting Whisper to low-resource language contexts.

Khade et al. (2025), in *Challenges in Adapting Multilingual LLMs to Low-Resource Languages using LoRA PEFT Tuning*, investigated the challenges of adapting Large Language Models (LLMs), specifically Gemma models, to Marathi, a low-resource language, using Low-Rank Adaptation (LoRA) Parameter-Efficient Fine-Tuning (PEFT). While automated evaluation metrics suggested a performance decline after fine-tuning, manual assessments revealed that the fine-tuned models often outperformed their original versions, particularly in generating contextually relevant responses.

4 Shared Task on Natural Language Understanding of Devanagari Script Languages

This shared task (Thapa et al., 2025b) aimed to address critical challenges in understanding Devanagari-script languages, which include Hindi, Nepali, Marathi, Sanskrit, and Bhojpuri. By focusing on language identification, hate speech detection, and target classification, the task sought to develop robust and generalizable NLP models for these linguistically rich but underrepresented languages. The task attracted widespread participation with 113 participants.

4.1 Shared Task Description

The shared task comprised three subtasks to explore different aspects of Devanagari-script language understanding. **Subtask A** focused on identifying the language of a given text among five Devanagari-script languages. **Subtask B** involved detecting whether a given text contained hate speech. **Subtask C** required identifying the target of hate speech, categorizing it as directed toward an individual, a community, or an organization. These tasks encouraged the development of models capable of addressing linguistic complexity and cultural nuances in diverse contexts.

4.2 Winning Team Performances

The winning teams employed innovative methodologies and domain-specific adaptations to achieve state-of-the-art performance across all three subtasks, showcasing the potential of multilingual and low-resource NLP research. Below, we give a short description of the approaches for winning teams in each subtask.

4.2.1 Subtask A

Team **CUFE** (Ibrahim, 2025) utilized fastText classifier for language identification, leveraging its subword modeling capabilities through n-grams along with systematic token generation using the tokenizer vy Team et al. (2022). The proposed system achieves a near-perfect F1 score of 0.9997 on the test set and secures the first position in the shared task.

4.2.2 Subtask B

Team **Paramananda** (Acharya et al., 2025) utilized FastText and demonstrated superior performance, particularly with data augmentation, achieving an F1 score of 81.39% and scoring first position on the leaderboard. This outperformed BERT, which struggled with an F1 score of 0.5763. Despite its contextual embedding strengths, BERT's underperformance was attributed to overfitting on sparse datasets, as evidenced by a higher evaluation score that did not generalize to test data.

4.2.3 Subtask C

Team **MDSBots** (Thapaliya et al., 2025) used a hybrid approach for the detection of the targets of

hate speech. Their approach involved augmenting the data with synthetic examples using synonym replacement and GPT-4, addressing class imbalance for minority categories like 'community'. Additionally, a rule-based Named Entity Recognition (NER) tagger was applied to identify entities such as individuals, organizations, and groups within tweets. These NER tags were incorporated into the model's input to improve performance, resulting in the highest F1 score in the competition. NER has historically shown good performance in target identification (Thapa et al., 2023).

5 Discussion on Challenges in Processing South Asian Languages

South Asian language processing poses significant challenges that stem from the region's linguistic diversity, complex scripts, and limited availability of resources. These challenges are compounded by the region's rich cultural and dialectal variations, creating additional obstacles for NLP applications. In this section, we examine several of these key challenges, drawing insights from accepted papers.

5.1 Low-Resource Nature and Data Scarcity

One of the foremost challenges is the lack of adequate linguistic resources, including annotated datasets, corpora, and pre-trained models for many South Asian languages. As highlighted by Thapa et al. (2025a), the scarcity of monolingual corpora and benchmarks limits the development of robust language models. For example, Nepali has limited large-scale datasets, and efforts such as pretraining language models on collected corpora are necessary to bridge this gap. Similarly, the lack of datasets for figurative language processing, such as idioms in Nepali (Pokharel and Agrawal, 2025), highlights the need for resource development to tackle specific linguistic phenomena.

5.2 Script and Orthographic Complexity

The South Asian linguistic landscape includes over 25 major scripts, many of which have complex orthographic rules that challenge standard encoding and rendering systems. Issues such as conjunct consonants, diacritical marks, and inconsistent Unicode implementation often lead to errors in text representation. As discussed by Ghimire et al. (2025), the Devanagari script, used in Nepali, Hindi, and Marathi, presents unique challenges for automatic speech recognition (ASR) and transcription due to its character-level complexities. Solutions like Rule-Based Character Constituency Loss (RBCCL) show promise in addressing transcription errors in Devanagari script (Ghimire et al., 2025).

5.3 Dialectal and Cultural Variations

South Asian languages often have multiple dialects with significant lexical and syntactic variations, making the development of universal models difficult. This diversity is further complicated by the lack of standardization in data collection and annotation. Papers like Chavinda and Thayasivam (2025) demonstrate how multilingual models must account for these variations to improve tasks like hate speech detection in south Asian languages like Tamil and Sinhala. Data augmentation and domain adaptation are critical for enhancing model performance across diverse dialectal contexts.

5.4 Linguistic Challenges

Languages like Tamil and Telugu, with their agglutinative morphology and intricate sandhi rules, challenge tokenization and translation systems. Specialized algorithms are required to address these linguistic features effectively (Dasari et al., 2025).

5.5 Code-Mixing and Multilinguality

Code-mixing, or the blending of languages within the same text, is a prevalent phenomenon in South Asia, particularly in social media and informal communication. This introduces additional challenges for NLP tasks like hate speech detection and sentiment analysis. The work by Srivastava (2025) exemplifies this issue by focusing on code-mixed Hindi-English hate speech detection. Fine-tuning multilingual models for such mixed-language contexts is an ongoing research challenge.

5.6 Speech and Text Processing

Speech recognition and text-to-speech systems face unique difficulties in South Asian languages due to phonetic richness and resource constraints. The lack of diverse speech datasets, as addressed by Chhetri and Poudyal (2025), limits the development of robust models. Similarly, the introduction of datasets like EmoTa for Tamil speech emotion recognition (Thevakumar et al., 2025) is a step forward in addressing the need for expanding resources in underrepresented languages.

5.7 Bias and Evaluation Limitations

The inherent biases in multilingual pre-trained models pose another significant challenge. Models trained on limited or skewed data often fail to generalize across different languages and tasks. Tahir et al. (2025) emphasize the need for benchmarking models on diverse set of tasks and languages, such as Urdu, to identify and mitigate these biases effectively.

5.8 Resource and Infrastructure Constraints

Unlike high-resource languages, South Asian languages suffer from limited computational and financial resources, which hampers the development of advanced models. The use of parameter-efficient fine-tuning methods like LoRA (Khade et al., 2025) offers a promising direction to address these constraints, enabling efficient model adaptation for low-resource settings.

5.9 Future Directions

While some progress has been made, addressing these challenges requires a multifaceted approach. Collaborative resource creation, the development of domain-specific models, and culturally informed annotation practices are critical for advancing NLP in South Asian languages. Furthermore, as demonstrated by CHiPSAL shared tasks (Thapa et al., 2025b), targeted competitions and benchmarks can drive innovation and foster community-driven solutions to these complex problems. In summary, addressing the challenges in processing South Asian languages requires a combination of innovative methods, resource development, and community collaboration. By tackling these issues, we can enhance inclusivity and robustness of NLP systems for the diverse linguistic landscape of South Asia.

6 Future Directions of the Workshop

In the future, the CHiPSAL workshop aims to expand its role as a leading platform for addressing the challenges and opportunities in South Asian language processing. Building upon the success of its inaugural workshop, the organizers plan to diversify the workshop's activities to foster deeper engagement and collaboration within the community. To encourage active participation and knowledge exchange, we aim to introduce interactive sessions such as expert panels, hands-on tutorials, and focused round-table discussions. These sessions will highlight emerging issues, including improving low-resource NLP methods, mitigating biases in multilingual and multimodal systems, and advancing cultural and linguistic inclusivity in AI.

Future workshops will place a strong emphasis on resource creation and open collaboration. We plan to organize dedicated tracks for dataset curation, multilingual benchmarking, and model development tailored to South Asian languages. These initiatives will address the lack of publicly available resources and benchmarks, empowering researchers and practitioners to develop state-of-the-art solutions for underrepresented languages. We also intend to expand the scope of shared tasks, introducing new challenges that encompass speech processing, code-mixing, figurative language understanding, and cross-modal applications. These tasks will encourage participants to tackle diverse linguistic phenomena and real-world scenarios unique to South Asia.

As NLP evolves, future CHiPSAL workshops will focus on leveraging advances in LLMs and multimodal systems for South Asian languages. This will include exploring innovative techniques such as continual learning, low-resource fine-tuning, and transfer learning, ensuring that state-of-the-art technologies are effectively adapted to the linguistic diversity of the region. Additionally, the workshop will continue to address ethical considerations, such as mitigating biases and ensuring fair representation in AI systems for South Asian languages. By promoting community-driven solutions and interdisciplinary collaboration, we aim to establish CHiPSAL as a key platform for the development of inclusive and impactful NLP in South Asia.

7 Conclusion

The inaugural CHiPSAL workshop provided a platform to address the challenges and opportunities in processing South Asian languages, emphasizing their linguistic diversity, script complexity, and low-resource nature. With contributions ranging from new datasets and benchmarks to advanced model fine-tuning, the workshop welcomed innovative approaches to tackle issues like code-mixing, dialectal variation, and linguistic resource constraints. Our shared task, which was participated by over 100 participants, also helped students and early-career researchers to understand and work on the problems within Devanagari NLU. Moving forward, CHiPSAL aims to expand its scope, foster

greater collaboration, and address emerging issues, with a focus on resource creation, ethical AI practices, and equitable access. By driving impactful research and fostering community engagement, CHiPSAL aspires to create lasting contributions that empower both academic and applied NLP for South Asia's rich linguistic and cultural heritage.

References

Darwin Acharya, Sundeep Dawadi, Shivram Saud, and Sunil Regmi. 2025. Paramananda@NLU of Devanagari Script Languages 2025: Detection of Language, Hate Speech and Targets using FastText and BERT. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia.

Krishan Chavinda and Uthayasanker Thayasivam. 2025. A Dual Contrastive Learning Framework for Enhanced Hate Speech Detection in Low-Resource Languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Ganesh Dhakal Chhetri and Prakash Poudyal. 2025. Impacts of Vocoder Selection on Tacotron-based Nepali Text-To-Speech Synthesis. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Priyanka Dasari, Mupparapu Sohan Gupta, Nagaraju Vuppala, Pruthwik Mishra, and Parameswari Krishnamurthy. 2025. Sandhi Splitting in Tamil and Telugu: A Sequence-to-Sequence Approach Leveraging Transformer Models. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Prakash Dhakal and Daya Sagar Baral. 2025. Abstractive Summarization of Low Resourced Nepali Language using Multilingual Transformers. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Sharad Duwal, Suraj Prasai, and Suresh Manandhar. 2025. Domain-Adaptive Continual Learning for Low-Resource Tasks: Evaluation on Nepali. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, 27th edition. SIL International, Dallas, TX, USA.

Rupak Raj Ghimire, Prakash Poudyal, and Bal Krishna Bal. 2025. Improving Accuracy of Low-resource ASR using Rule-Based Character Constituency Loss (RBCCL). In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

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

Michael Ibrahim. 2025. CUFE@NLU of Devanagari Script Languages 2025: Language Identification using fastText. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Antony Alexander James and Parameswari Krishnamurthy. 2025. POS-Aware Neural Approaches for Word Alignment in Dravidian Languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

A H M Rezaul Karim and Ozlem Uzuner. 2025. Leveraging machine-generated data for joint intent detection and slot filling in bangla: A resource-efficient approach. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2025. Challenges in adapting multilingual llms to low-resource languages using lora peft tuning. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Ayesha Khalid, Farah Adeeba, Najm Ul Sehar, and Sarmad Hussain. 2025. Bridging the Bandwidth Gap: A Mixed Band Telephonic Urdu ASR Approach with Domain Adaptation for Banking Applications. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Aayush Neupane, Aayush Lamichhane, Ankit Paudel, and Aman Shakya. 2025. Structured Information Extraction from Nepali Scanned Documents using Layout Transformer and LLMs. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Rhitabrat Pokharel and Ameeta Agrawal. 2025. neDIOM: Dataset and Analysis of Nepali Idioms. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

P Rahothvarman, Adith John Rajeev, Kaveri Anuranjana, and Radhika Mamidi. 2025. Bridge the GAP: Multi-lingual Models For Ambiguous Pronominal Coreference Resolution in South Asian Languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking whisper for low-resource speech recognition: An n-shot evaluation on pashto, punjabi, and urdu. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Varad Srivastava. 2025. DweshVaani: An LLM for Detecting Religious Hate Speech in Code-Mixed Hindi-English. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking the Performance of Pre-trained LLMs across Urdu NLP Tasks. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Nahida Akter Tanjila, Afrin Sultana Poushi, Sazid Abdullah Farhan, Abu Raihan Mostofa Kamal, Md. Azam Hossain, and Md. Hamjajul Ashmafee. 2025. Bengali ChartSumm: A Benchmark Dataset and Study on Feasibility of Large Language Models on Bengali Chart to Text Summarization. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Prajwal Thapa, Jinu Nyachhyon, Mridul Sharma, and Bal Krishna Bal. 2025a. Development of Pre-Trained Transformer-based Models for the Nepali Language. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 2023. Multimodal Hate Speech Event Detection-Shared Task 4. In *CASE 2023-Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 151–159. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025b. Natural language understanding of devanagari script languages: Language identification, hate speech and its target detection. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Anish Thapaliya, Prabhat Ale, and Suman Paudel. 2025. MDSBots@NLU of Devanagari Script Languages 2025: Detection of Language, Hate Speech, and Targets using MURTweet. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Uthayasanker Thayasivam, Thulasithan Gnanenthiram, Shamila Jeewantha, and Upeksha Jayawickrama. 2025. SiTa - Sinhala and Tamil Speaker Diarization Dataset in the Wild. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).

Jubeerathan Thevakumar, Luxshan Thavarasa, Thanikan Sivatheepan, Sajeev Kugarajah, and Uthayasanker Thayasivam. 2025. EmoTa: A Tamil Emotional Speech Dataset. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi, UAE. International Committee on Computational Linguistics (ICCL).