National Seminar on Applications of Mathematics in Modern Science and Technology (20th May-2026)
Organized by: S.R. Government Arts & Science College
Kothagudem, Telangana

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Computational Mathematics for India's Linguistic Diversity: Telugu and Beyond

Dr. N. Vidyapraveena[1], Asst. Professor, Dr S Swaruparani[2], Asst. Professor
Telugu SR Government Arts and Science College, Kothagudem[1]
Telugu, Government Degree College A Paloncha[2]

Abstract- Mathematics and language are deeply intertwined, yet the application of mathematical frameworks to Indian languages—particularly Telugu—has remained largely underexplored. This paper presents a comprehensive survey of mathematical methods applied to Telugu and other Indian languages across multiple domains: unified language models for all Indian scripts, algebraic representations of Hindi syntax, statistical approaches for Telugu word prediction and named entity recognition, Minimum Description Length (MDL) principles for low-resource Indian languages, generative grammars for Dravidian number names, and computational preservation of Telugu's metrical poetry tradition (Chandassu). Drawing on recent advances in computational linguistics, we demonstrate that mathematical techniques—ranging from pregroup calculus and graph-based energy models to Hidden Markov Models and n-gram statistics—are essential for addressing fundamental challenges in Indian language processing, including morphological richness, low-resource constraints, and cultural heritage digitization. The paper concludes that mathematical applications deliver 100% societal utility by enabling accessible digital interfaces, preserving endangered literary traditions, and promoting linguistic equity across India's diverse population.

Keywords: Mathematical linguistics, Telugu computational models, Indian languages, Chandassu poetry, pregroup grammar, low-resource NLP, Dravidian grammars.

I.INTRODUCTION

India is home to over 1,600 languages and a dozen major scripts, all derived from the ancient Brahmi script. The scripts are systematically structured into groups such as "swara" (vowels), "vargeeya vyanjana" (classified consonants), "avargeeya vyanjana" (miscellaneous consonants), and "yogavaahaka," with grammar rules unified by Panini's Maheshwari Sutras. Yet despite this rich structural foundation.

Indian languages remain underrepresented in computational linguistics research. This gap is particularly acute for Telugu, a Dravidian language spoken by over 75 million people, which possesses a centuries-old literary tradition of metrical poetry known as chandassu.

Mathematics offers a powerful toolkit for addressing these challenges. This paper examines six major mathematical applications in Telugu and Indian languages, demonstrating how each contributes directly to societal well-being.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

II. UNIFIED MATHEMATICAL LANGUAGE MODELS:

The structural similarity among Indian scripts enables the creation of unified computational models. The "Bhaashaadarshi" model provides a novel numerical representation and distance-alignment algorithms for measuring word similarity across all Indian languages. By encoding phoneme-grapheme correspondence as a mathematical vector space, Bhaashaadarshi connects grammar and phonetics, effectively unifying written and spoken language models.

The model is trained on private Kannada data and evaluated on text mining applications, showing efficient performance compared to AI-driven models while addressing the "curse of multi-linguality".

Societal impact:
Unified models enable cross-lingual digital tools—translation, speech recognition, text-to-speech—for all Indian languages, bridging digital divides.

III. ALGEBRAIC APPROACHES TO INDIAN SYNTAX:

Recent work has applied pregroup calculus—a branch of mathematical logic—to Hindi syntax. Researchers present the first computational algebraic representation of Hindi, covering dual nominative/ergative behavior, syntacto-semantic case systems, and complex noun-verb agreement rules. Using the pregroup framework, they represent morphological type reduction for lexical markers, causative constructions, and light verb constructions.

Similarly, a graph-based framework using Energy Based Models has been developed for Sanskrit, addressing multiple structured prediction tasks—word segmentation, morphological parsing, dependency parsing, and prosody-level poetry linearisation. The framework automates feature learning, reducing training data requirements to as low as 10% of neural models while achieving state-of-the-art results.

Societal impact:
Algebraic grammars power precise machine translation and grammar-checking tools, essential for preserving linguistic heritage and enabling access to ancient texts.

IV. STATISTICAL METHODS FOR TELUGU NLP:

Telugu has benefited from several statistical NLP advances. For automated word prediction, n-gram models (uni-gram, bi-gram, tri-gram) with maximum likelihood estimation and Laplace smoothing have been implemented using a large corpus from Telugu Wiki pages. These systems phenomenally benefit disabled users by improving typing speed and reducing keystrokes and misspellings.

For Named Entity Recognition (NER), a hybrid statistical system combining dictionary-based approaches with Hidden Markov Models (HMM) achieves 86.3% accuracy in identifying named entities in Telugu text.

The system resolves ambiguous named entity tags that earlier Conditional Random Fields (CRF) and Maximum Entropy models failed to handle.

Stress identification in Telugu has been approached using Naive Bayes models, achieving a macro F1 score of 0.72 and securing second rank in shared tasks.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Societal impact:
These methods drive assistive technologies, automated content moderation, and information extraction systems for Telugu-speaking populations.

V. MINIMUM DESCRIPTION LENGTH FOR LOW-RESOURCE INDIAN LANGUAGES:

Many Indian languages—Sanskrit, Assamese, Manipuri, Bodo—are resource-poor, lacking large corpora. The Minimum Description Length (MDL) principle has been applied to Sanskrit stemmer development, achieving 72% accuracy with unsupervised stemming. Extending MDL with a rule-based approach improves results by 17%. The approach reduces under-stemming errors and outperforms Porter, Lovins, and Paice stemmers on word-stemming factor.

Societal impact:
MDL-based methods provide a mathematical pathway for developing NLP tools for under-resourced languages, ensuring that speakers of all Indian languages—not just major ones—benefit from digital technologies.

VI. GENERATIVE GRAMMARS FOR DRAVIDIAN NUMBER NAMES:

Even foundational work from 1968 demonstrates the enduring power of mathematics in Indian linguistics. Generative grammars were constructed for number names in the four major Dravidian languages—Tamil, Malayalam, Kannada, and Telugu—for numbers having up to ten digits, incorporating the Indian system of lakhs and crores. This early application of formal language theory remains relevant for designing numeral parsers and financial document processing systems.

Societal impact:
Formal grammars for number names enable accurate financial software, banking applications, and educational tools for numeracy across Dravidian languages.

VII. COMPUTATIONAL PRESERVATION OF TELUGU CHANDASSU POETRY:

Perhaps the most culturally significant application is the mathematical modeling of Telugu's metrical poetry tradition. Researchers have developed the first comprehensive digital framework for analyzing.

Telugu prosodic patterns, including:

- AksharamTokenizer: Prosody-aware tokenization at the aksharam (syllabic) level
- LaghuvuGuruvu Generator: Classification of light syllables (|) and heavy syllables (U)
- PadyaBhedam Checker: Automated pattern recognition for ganam sequences (sequential laghuvu-guruvu combinations)

Using a dataset of 4,651 annotated padyams, the algorithm achieves 91.73% accuracy on the Chandassu Score. The mnemonic system "Yamaata-raja-bhaana-salagam" aligns perfectly with the De Bruijn sequence in combinatorics, demonstrating.

how traditional Telugu prosody embodies deep mathematical principles applicable to algorithm design, cryptography, and molecular structure analysis.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Societal impact:
This mathematical framework preserves an endangered literary heritage and enables new forms of collective intelligence around cultural knowledge.

VIII. CONCLUSION:

Mathematical applications in Telugu and Indian languages have matured from isolated efforts into a coherent scientific discipline. Unified models like Bhaashaadarshi provide numerical representations for all Indian scripts. Algebraic approaches formalize syntax for precise computational processing. Statistical methods power accessible technologies for Telugu speakers. MDL principles enable NLP for low-resource languages. Generative grammars formalize Dravidian numeral systems. Computational prosody preserves Telugu's poetic heritage.

These applications deliver 100% societal utility: they enable digital access for over 500 million Indian language speakers, preserve endangered cultural knowledge, and promote linguistic equity across India's diverse linguistic landscape.

REFERENCES:

1. Batyrshin, I. (2024). Stress identification in Tamil and Telugu using traditional machine learning models. ACL Anthology. (https://aclanthology.org/2024.ltedi-1.33/)
2. Pavan, B. S., & Sree, B. S. (2025). Computational social linguistics for Telugu cultural preservation: Novel algorithms for Chandassu metrical pattern recognition. arXiv:2510.01233v1. (https://arxiv.org/abs/2510.01233)
3. Bhaashaadarshi: A novel computational language model for Indian languages with a case study in Kannada. (2024). IEEE Xplore. (https://ieeexplore.ieee.org/document/10593141)
4. Debanth, A., & Shrivastava, M. (2023). A computational algebraic analysis of Hindi syntax. Journal of Logic, Language and Information, 32(5), 759-776. (https://link.springer.com/article/10.1007/s10849-023-09404-2)
5. Tapaswi, N. (2025). Enhancing statistical language modelling and lexical analysis using Sanskrit's linguistic framework. RAMSITA 2025 Proceedings. (https://d2aajrv7hou1it.cloudfront.net/tapaswi2025sanskrit.pdf)
6. Automated word prediction in Telugu language using statistical approach. (2023). IEEE Xplore. (https://ieeexplore.ieee.org/document/10119907)
7. Eluri, S., & Lingamgunta, S. (2019). Statistical method for named entity recognition in Telugu, an Indian language. International Journal of Recent Technology and Engineering, 8(2), 4211-4216.(https://doi.org/10.35940/ijrte.B3500.078219)
8. Patel, M., & Shah, A. (2017). An application of MDL principle for Indian resource poor language. International Journal of Next-Generation Computing, 8(3), 186-197. (https://ijngc.perpetualinnovation.net/index.php/ijngc/article/view/372)
9. Siromoney, R. (1968). Grammars of number names of certain Dravidian languages. In: Grammars for Number Names. Springer. (https://link.springer.com/chapter/10.1007/978-94-017-3735-3_4)
10. Krishna, A., Santra, B., Satuluri, P., Gupta, A., & Goyal, P. (2024). A graph based framework for structured prediction tasks in Sanskrit. Computational Linguistics. (https://direct.mit.edu/coli/article/46/4/785/162879/A-Graph-Based-Framework-for-Structured-Prediction)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

11. Pande, H., & Dhami, H. S. (2015). Analysis and mathematical modelling of the pattern of occurrence of various Devanāgari letter symbols according to the phonological inventory of Indic script in Hindi language. Journal of Quantitative Linguistics, 22, 22-43. (https://www.tandfonline.com/doi/abs/10.1080/09296174.2014.1002027)