Anon Author
University of Anon
Anonville, Rep. of Anon
anon@anonuni.edu

Anon Author
University of Anon
Anonville, Rep. of Anon
anon@anonuni.edu

Anon Author
University of Anon
Anonville, Rep. of Anon
anon@anonuni.edu

Anon Author
University of Anon
Anonville, Rep. of Anon
anon@anonuni.edu

ABSTRACT

Background:
Objective:
Method:
Results:

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; • **Computing methodologies** → **Artificial intelligence**.

ACM Reference Format:
Anon Author, Anon Author, Anon Author, and Anon Author. 2024. . In *Proceedings of the 2024 Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024), July 8–10, 2024, Milan, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

Bhattacharyya et al. [1] characterized the challenges of NLP in Indic languages as,

(1) **Scale and Diversity** Indic languages encompass a vast array of languages and dialects, belonging to multiple linguistic families and written in numerous distinct scripts.
(2) **Longer Utterances** Sentences in Indic languages are often longer and more complex than in English, complicating tasks like parsing and speech recognition.
(3) **Code Mixing** The frequent mixing of multiple languages in a single sentence or conversation is a common challenge in computational linguistics for the region.
(4) **Resource Scarcity** Many Indic languages lack sufficient annotated datasets for building robust NLP and speech tools.
(5) **Absence of basic speech and NLP tools** Foundational tools like morphology analyzers and speech recognition systems are either unavailable or lack accuracy for most Indic languages.
(6) **Absence of linguistics knowledge** A limited understanding of the linguistic structure of many regional languages hinders the development of computational models.
(7) **Script complexity and non-standard input mechanisms** The diversity of scripts and their associated vowel and consonant combinations make input systems slower and less intuitive.
(8) **Non-standard transliteration** Roman transliteration of Indic languages lacks standardization, leading to multiple ways of representing the same word.
(9) **Non-standard storage** Variations in how characters are encoded and stored pose issues in data sharing and tool interoperability.
(10) **Man-made Problems** Government-imposed standard keyboards and inadequate funding often stifle innovation and efficiency in linguistic computing.
(11) **Some challenging language phenomena** Features like free word order, agglutination, and context-dependent pronunciation introduce additional computational hurdles.

In Jordan et al. [2], they found that amoung the languages under their consideration, English, Spanish, Vietnamese, and Tamil, the worst performing language was Tamil, generating incorrect solutions and non-sensiable translations. Though the authors did not systematically investigate the the nature of the poor performance or methods of improvement, they did note several possible reasons:

(1)
(2) Compared to English and Spanish it was likely a lower resourced language in the training data.

The first of these two possibilities is, perhaps, the most interesting. In Tamil, there are two core dialects which differ significantly: literary Tamil (sen-Tamil) and colloquial (kodun-Tamil) []. Additionally, given Tamil is spoken not just in and around Tamil Nadu, but also in Sri Lanka, Malaysia, and Singapore, there are a number of regional dialects [].

2 BACKGROUND

2.1 Indic LLMs and Datasets

Ramesh et al. [3] introduced Samanantar, a large-scale multilingual dataset for Indic languages containing sentence pairs in 11 languages.

ITiCSE 2024, July 8–10, 2024, Milan, Italy

ACM ISBN 979-8-4007-XXXX-X/XX/XX
https://doi.org/10.1145/XXXXXX.XXXXXX

3 METHODS

4 RESULTS

5 DISCUSSION

6 LIMITATIONS

7 CONCLUSION

REFERENCES

[1] Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga, and Ranjiva Munasinghe. 2019. Indic language computing. *Commun. ACM* 62, 11 (2019), 70–75.

[2] Mollie Jordan, Kevin Ly, and Adalbert Gerald Soosai Raj. 2024. Need a Programming Exercise Generated in Your Native Language? ChatGPT's Got Your Back: Automatic Generation of Non-English Programming Exercises Using OpenAI GPT-3.5. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 618–624.

[3] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics* 10 (2022), 145–162.