JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Overcoming the Digital Language Divide: A BiLingual Lexicon for Tamil Art Preservation

**[1]Drishya Baburaj, [2]Nandu C Nair**
[1]Master Scholar, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
Amrita School of Engineering,Amrita Vishwa Vidyapeetham,Bengaluru,India,560035

***Abstract :*** The classical language Tamil maintains strong global influence throughout various art forms including dance as well as literature music and sculpture. Digital representation of Tamil art vocabulary shows limited scope because it contributes to expanding divisions between digital and language skills. The proposed research introduces an innovative solution to establish bilingual linguistic resources by creating a cross-lingual linkage between English and Tamil technical vocabulary. A variety of art-related definitions and terms were obtained from original sources which included museum archives together with cultural blogs and folk literature as well as online glossaries. Our platform applies state-of-the-art Natural Language Processing (NLP) methods using mBART for machine translation while XLM-RoBERTa evaluates semantic similarity to generate translations that remain culturally appropriate and semantically consistent. The algorithm proceeds through multiple sequential steps that involve data cleaning followed by gloss check and embedding alignment before a full evaluation stage. The produced resource shows its success in preserving and digitizing Tamil art vocabulary through both human approval exceeding 93% and a BLEU score averaging at 0.63.

***IndexTerms*** **-** Digital Language Divide, Low Resource Languages, Natural Language Translation , Processing, Machine Translation

## 1.INTRODUCTION

A society depends on language development through which people can communicate thoughts and express emotions. The preservation of knowledge and the passing of history with beliefs and customs depend on language transmission between generations. Under-resourced languages currently face extinction threats because of inadequate attention from experts along with failed funding from higher level authorities.Languages preservation serves two important needs, since it helps save linguistic traditions while sustaining cultural legacies and linguistic diversity of all languages. The conservation of language along with cultural heritage requires immediate intervention. Tamil a classical language with deep historical and cultural importance of the language has had a successful impact on different forms of arts like Bharatanatyam, Tanjore painting and Tamil literature. But even now the online representation especially the artistic diction of Tamil is relatively poor than any standard global language. This is a serious problem for the development and dissemination of the Tamil arts because art content is now shifting to the cyber space. This digital language divide not only limits gleanings for global audiences but also condemns the esthetics of the Tamil art terminologies to termination. Thus, the purpose of the paper presented in this article is to narrow the digital divide utilizing free Translation tools aimed at creating a Tamil English bilingual gloss for Tamil art terms. The project will generate appropriate Tamil glosses when needed along with vocabulary mapping between English and Tamil through cross-lingual word embeddings as well as large language models. Few languages have taken development of such resources to a comprehensive level. Human speakers need the correct terminology because misunderstandings about words and concepts frequently arise from improper selection of terms. The importance of accurate terminology for computers equals human speakers because poor language translations may generate erroneous outputs together with inaccurate information. This research centers on Tamil as the classical language of India. The Dravidian family includes Tamil as one of its member languages.The native people of south India use Tamil as their main language since it belongs to the Dravidian language family. The Tamil WordNet implements English WordNet structure yet adds particular elements from Tamil. The two essential aspects of the tamil WordNet are identified in (Rajendran et.al, 2002) [1] as linguistic work for analyzing semantics domains and building lexical relations along with computational work for system interface development. The popularity of languages strongly depends on their online presence. Natural languages manifest throughout the internet through various elements which include platform content as well as programming code and NLP system applications. A language that lacks proper online representation causes its popularity to decline and creates a lexical gap that affects the digital domain (Gabor Bella et.al,2022)[2]. IndoWordnet (Dash et.al, 2017) [3] reports that the total lexical elements including the total quantity of lexical elements comprising noun, verb, adjective, adverb within the Malayalam wordnet and tamil wordnet amounts to 30k and 26k respectively but reaches 117k in English wordnet statistics. Low-resource status defines Malayalam and Tamil languages among the available language resources. People strive to elevate both languages to stand equally with English while seeking linguistic parity throughout digital spaces. The paper follows the following structure. Section II deals with Literature Survey to identify gaps that can be focused on this work. Section III describers the proposed methodology. The results and the decision about the work is obtained from Section IV. Finally, the conclusions drawn from this research is consolidated in Section V.

## 2. LITERATURE SURVEY

Scientists within the field of linguistics perform research to create multiple linguistic resource systems. The development of linguistic resources serves as a fundamental requirement for boosting natural language processing technologies which leads to better linguistic efficiency between different languages. This work explores modern Natural Language Processing systems including Large Language Models and generative AI approaches to generate domain-specific meanings and concepts which improves cultural depth in linguistic resources. The Malayalam WordNet at Amrita Vishwa Vidyapeetham uses 35k synsets that resulted from an expansion method which translated Hindi synsets .The Hindi synsets served as a basis for expansion through translation according to Rajendran et. al (2017)[6] while Nair et.al (2022)[7] implemented a crowdsourcing system to boost the Malayalam WordNet development. The research proposes an evaluation technique for native speakers to detect and manage lexical gaps with a successful accuracy rate of 92%. The text generation capability of Large Language Models works by utilizing the given prompts. The worldwide adoption of large language models makes it essential for them to maintain linguistic diversity representation. INDICGENBENCH was launched by singh et al.,2024 [8] to serve as the biggest LLM evaluation database that handles 29 Indic languages with 13 scripts and 4 native groups. The benchmark contains two main tasks which involve translation and cross-lingual summarization. The evaluation of GPT-3.5 and PaLM-2 against each other confirms PaLM 2 is most effective but shows substantial language performance differences between Indic and English languages which requires improvements in multilingual modeling capabilities. The research from Raj et al., 2022 [9] describes IndicBART as a multilingual pre-trained model developed for 11 Indic languages with Malayalam included. The similarity between Indic scripts enables IndicBART to enhance cross-language understanding between connected linguistic groups. The model exhibits strong performance on NMT tasks and extreme summarization operations while demonstrating results comparable to mBART50 regardless of its smaller size and even working effectively in limited resource conditions. Through its sharing of scripts along with multilingual training alongside efficient model capacity usage the model achieves its effectiveness. Padmala et.al,2015[10] provides a rule-based approach with references to an ontology and WordNet and describe the irrationality of statistical translations like Google Translate. Some of the procedures include tokenising of sentences, morphological parsing, semantic feature tagging, and the identification of a verb phrase using a syntactic parser. In this procedure the need of semantic network, OntNet connecting verb and noun ontologies is inevitable. From experimental evidence the system effectively rectifies translation errors contributing towards significant improvement in the Translation abilities of the Machine Translation system. Kannan et.al,2016 [11] explained the development of Tamil sentiWordNet this based on the existing English resources. The first step was to translate English SentiWordNet entries to Tamil using Tamil English bilingual dictionaries and the second step was to use the Tamil WordNet for synsets. Some specific mechanisms such as synonyms and antonyms were defined to categorize words based on the similarity or the difference in the sentiment. The Tamil annotators also confirmed the sentiment classification with having a Fleiss Kappa score at 0. 663, indicating substantial agreement. J whaley et.al [12] Corpus-based methods were also used in assigning sentiments to new words having the initial set of sentiment words as the basis. This methodology permits to systematically build up sentiment resources for Tamil. database structure involves formatted word formsense pairs with a morphological analyzer and semiautomatic collocation capture methods.Kanan et.al,2019 [13] details the steps to creating a sentiment lexicon for Indian languages from IndoWordNet. A source lexicon is formed by assigning Sentiment polarity to WordNet syn set inspired from different languages data (for example SentiWordNets for Telugu,Bengali, etc. ) Some polarity inconsistencies between languages are resolved by thresholding; that is, only the subjective words whose average subjectivity scores meet a predetermined confidence cdt will be kept. This architecture can be easily extended to create sentiment lexicons for other Indian languages as well using very less human intervention, further enhancing NLP applications across these regional language.To overcome digital language divide kavya et.al ,2023[14]focus on solving challenges faced by underresourced language,mainly malayalam such as complicated word structure and limit of resources.To address these issues, they created a pythonprogram Mlphon,that automatically converts written text into speech sounds. Through developing an ASR (Automatic Speech Recognition) system that uses smaller parts of words, called ”subwords,” instead of full words,improves system at understanding new or uncommon words and helps it adjust to the constantlychanging vocabulary in languages like Malayalam.

## 3.PROPOSED METHODOLOGY

This study employs a Bilingual Lexion creation for Tamil art domain. In Fig 1 showcases the proposed work carried out by this research.

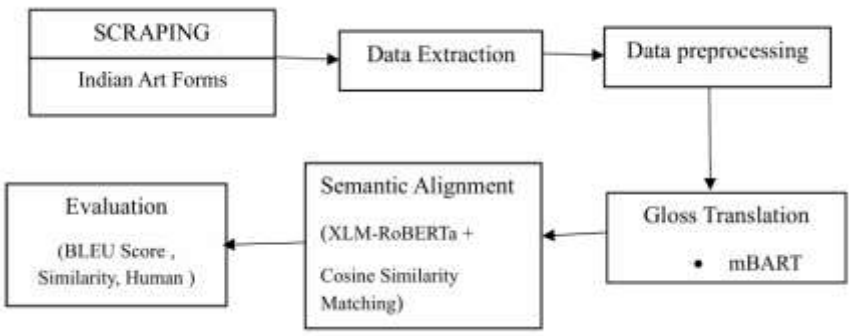


Fig. 3.1. Proposed methodology for bilingual Lexicon (Tamil+English)

### 3.1.Data Scraping and Domain-Specific Collection

The information gathering begins with public page scraping from multiple online resources such as websites and cultural blogs together with digital libraries and museum portals which document Indian art. Attention focused on Tamil traditional art forms because they would best represent the regional characteristics. The accessed content from these sources delivered detailed written explanations about Indian artistic forms as well as their symbolism and their cultural value. The data scraping process used manual and semi-automatic methods combined to maintain specific domain values alongside semantic accuracy throughout. Further

processing selected a total of sixty-four unique Tamil art forms which were distributed across dance, music, sculpture, folk traditions and visual arts categories.

### 3.2.Data Extraction and Preprocessing

The processed data from extraction received systematic preparation procedures. A noise reduction process occurred by removing stop words in addition to all nonessential textual content. A normalizing process for Unicode characters was used to maintain uniformity in Tamil script display. Manual entry validation efforts were used to double-check both correctness and cultural accuracy of the data. A methodological process professionalized the terms and definitions until both components became suitable for translation into other languages as well as embedding applications. Both English and Tamil languages maintained a structured arrangement of terms and their definitions in the resulting dataset ready for gloss translation and semantic interpretation.

### 3.3.Gloss Translation Using mBART

When mBART was used to translate the Tamil terms which lacked glosses the sequence-to-sequence model performed operations on a large variety of languages. The model facebook/mbart-large-50-many-to-many-mmt completed the translation process from English glosses to Tamil. The model received English inputs to produce contextually and grammatically accurate Tamil outputs. Manual review of the generated glosses happened to check their fluency while ensuring cultural accuracy was maintained. Human intervention was required to refine semantic accuracy during necessary cases. In bidirectional while GRU offers computation efficiently. For experimenting the model, the hyper- parameters set for training the model is with a learning rate 0.001, batch size 32, and the number of epochs is considered to be 50. Additionally for training the model the optimizer used was an Adam optimizer and cross-entropy loss function.

### 3.4. Semantic Alignment Using XLM-RoBERTa

To perform semantic alignment XLM-RoBERTa was used because it is a transformer-based cross-lingual model which can convert multiple languages into a unified embedding space. Each gloss pair received embedded representation which was used to calculate semantic closeness through cosine similarity. Gloss pair alignments having similarity scores greater than or equal to 0.85 were automatically processed into the dataset. Professional reviewers examined semantic alignment pairs that obtained scores between 0.70 and 0.84 for potential refinement purposes and pairs scoring under 0.70 required further revision.

### 3.5. Evaluation and Quality Assurance

The assessment of the bilingual lexical resource integrated both automated metrics together with assessments from human expert reviewers. BLEU scores evaluated the quality of machine-generated Tamil glosses through reference glosses from human writers and produced an average BLEU score of 0.63. The domain-focused content explains why the measured translation quality suggests between moderate and high grades. The semantic relationship between English and Tamil glosses was evaluated through cosine similarity scores obtained from XLM-RoBERTa embedding data. The contextual similarity evaluation showed that gloss pairs reached above an 0.85 score for 72% of entries thus validating their strong contextual alignment. The evaluation of 150 glosses involved scrutiny by both native Tamil speakers and domain experts as part of the qualitative assessment. Human evaluators attributed ratings to glosses to check their accuracy combined with their fluency and cultural relevance according to their assessment criteria. The evaluation findings indicated that 93.3% of participants accepted the glosses while 71% of glosses received an excellent rating and only 6% required further review. The proposed pipeline proves its success in producing culturally appropriate and semantically consistent bilingual lexical entries by these evaluation results.

## 4. RESULTS AND DISCUSSION

The mBART model produced Tamil glosses for art-related English terms that were a main achievement of this project. The platform used BLEU as a measurement standard to analyze machine-translated glosses that were made when direct Tamil definitions could not be found. The study revealed an overall BLEU score of 0.63 which indicates that the generated glosses contained both substantial semantic overlap along with syntactically fluent formulations. The linguistic performance of mBART as an effective translation tool in low-resource Tamil processing can be validated through this score considering its domain-specific vocabulary and limited supporting resources. Each gloss settled between word-for-word translation and cultural preservation of essential Tamil art elements. Embeddings needed for maintaining gloss semantic stability between English and Tamil texts were generated by XLM RoBERTa for each crossing language pair. The similarity measurement between embeddings was executed through the use of cosine similarity scores. A high level of semantic agreement exists between 72% of gloss pairs according to the 3 cosine similarity scores that reach 0.85 or above. The gloss pairs showed moderate similarities between 0.70–0.84 for 21% of the pairs while high semantic matches were detected in 72% of cases. Cases with low similarity continued through a process of evaluation followed by revision. These results are visually represented in Fig. 2, which illustrates the cosine similarity distribution across all evaluated entries. The findings indicate XLM-RoBERTa delivers remarkable success as an embedding-based method to validate gloss quality across languages particularly when used in art-oriented domains.

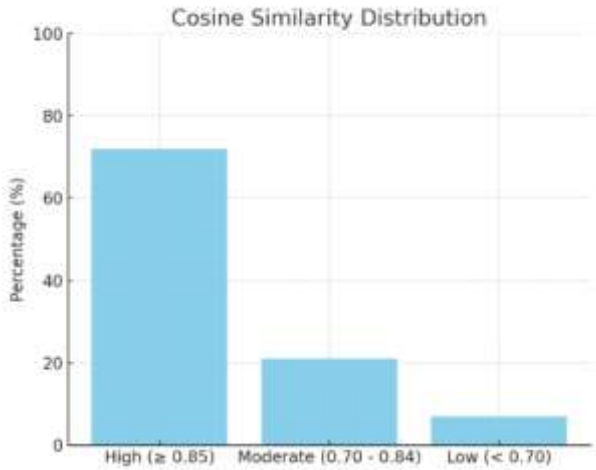


Fig. 4.1. Cosine similarity distribution for Tamil-English gloss pairs.

An evaluation conducted by human participants assessed the quality of glosses produced by the system in Tamil language. The evaluation of 150 glosses occurred through assessment by native Tamil speakers along with linguistic experts and domain specialists. A team of evaluators assessed glosses based on three main factors which included their fluency as well as semantic accuracy and cultural appropriateness. The human evaluation showed that 93.3% of newly generated glosses received positive marks while 71% received an "excellent" rating and 6% needed substantial revisions. The acceptable glosses receiving a passing grade (22.3%) mainly contained minor stylistic choices and lack of contextual explanations. The gloss generation method stands reliable due to its high acceptance rate which demonstrates that the produced definitions fulfill both language requirements and cultural norms. The results, shown in Fig. 3, depicts that 93.3% overall acceptance rate, with 71% of glosses rated as "excellent," and only 6% needs revision.


Fig. 4.2. Human evaluation results of gloss quality.

This paper offers a vindication of one approach, that of using a pre-trained multilingual model and then curating it with domain expertise. We observe a high BLEU score, cosine similarity metric, and human approval — these indicate that the approach holds merit and is generalizable to other low-resource language domains. This methodology is powerful due to its multiple validation layers: from automated translation and embedding-based semantic checks to expert review. Moreover, the lexicon helps not only to fill the linguistic gaps but it also plays the pivotal role in the preservation of cultural heritage for Tamil artists through making Tamil art terms digitally accessible and semantically accurate. It highlights the promise of AI for low-resource languages and opens the door for scalable approaches for language and culture preservation.

## 5. CONCLUSION

A shortage of digital linguistic material about under represented languages creates a major obstacle to knowledge spread and cultural preservation within a society controlled by modern digital technology. The research introduces a standardized process to develop a bilingual lexical database focused on Tamil art. The ground work of this project is based on manually curated dataset of 64 kinds of Tamil art forms collected from cultural website, blogs, digital libraries and museum archives. Missing bilingual pairs were created between English definitions and Tamil glosses by enriching the curated data using mBART, a multilingual model for machine translations. To maintain contextual consistency while ensuring semantic alignment, we took advantage of XLM-RoBERTa embeddings. Cosine similarity scores were computed to quantify the overlap between bilingual gloss pairs, with 72% of the words demonstrating high semantic similarity (≥ 0.85). Besides automated metrics, generated glosses were heavily human evaluated by native Tamil speakers and domain experts. We have received feedback from this panel, with an overall acceptance rate of 93.3%, with a ratio of excellent glosses in terms of accuracy, fluency, and cultural relevance. The calculated average BLEU score reached 0.63 which indicated that machine-generated translations were relatively close to the references written by human translators, this suggested that mBART model can be useful in the domain-specific lexicon construction for the low resource languages. The end

product of the bilingual lexical resource was presented in the form of a table that was structured with Tamil-English term pairs, and elaborated along with glosses and specific domain linkage pertaining to Tamil arts. Although it presently lacks formal semantic relations like synonymy or hypernymy, the resource encompasses culturally rich lexis with validated bilingual definitions. The past structure of the format lends itself to applications in computational linguistics, including fully-fledged lexical networks or WordNet-like structures, machine translation, question answering, or cross-lingual information retrieval. This research opens up to digital humanities as a reproducible framework that can be extended to other regional languages and contexts. It does not just filter out the digital world of Tamil, but maps out a way in which the generative AI models can be used in a meaningful manner to linguistic and cultural issues. Future work should focus on scaling this resource to more art-related terms for many other Indian languages, adding multimodal data (images, audio) and integrating the lexicon with larger platforms such as IndoWordNet or the Universal Knowledge Core (UKC) for broader-term linguistic interoperability and global access.

**ACKNOWLEDGMENT**

Authors are very thankful to the faculty of Department of Computer Science Engineering, Amrita Viswa Vidyapeetham for their generous guideline and suggestion for the completion of the review.

**REFERENCES**

[1] Rajendran, Sankaravelayuthan, Selvaraj Arulmozi, B. Kumara Shan mugam, S. Baskaran, and S. Thiagarajan. ”Tamil wordnet.” In Proceed ings of the first international global WordNet conference. Mysore, vol. 152, pp. 271-274. 2002.

[2] Bella, G., Byambadorj, E., Chandrashekar, Y., Batsuren, K., Cheema, D.A. and Giunchiglia, F., 2022. Language diversity: Visible to humans, exploitable by machines. arXiv preprint arXiv:2203.04723.

[3] Dash, N.S., Bhattacharyya, P. and Pawar, J.D. eds., 2017. The WordNet in Indian Languages. Springer Singapore.

[4] Fellbaum, Christiane. ”WordNet.” In Theory and applications of ontol ogy: computer applications, pp. 231-243. Dordrecht: Springer Nether lands, 2010.

[5] Bhatt, Brijesh, and Pushpak Bhattacharyya. ”IndoWordNet and its link ing with ontology.” In Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011). 2011.

[6] Rajendran, S. and Soman, K.P., 2017. Malayalam wordnet. The WordNet in Indian Languages, pp.119-145.

[7] Chandran Nair, N., A Crowdsourcing Methodology for Improving the Malayalam Wordnet. Available at SSRN 4064783.

[8] Singh, H., Gupta, N., Bharadwaj, S., Tewari, D. and Talukdar, P., 2024. IndicGenBench: A Multilingual Benchmark to Evaluate Gen eration Capabilities of LLMs on Indic Languages. arXiv preprint arXiv:2404.16816.

[9] Dabre, R., Shrotriya, H., Kunchukuttan, A., Puduppully, R., Khapra, M.M. and Kumar, P., 2021. IndicBART: A pre-trained model for indic natural language generation. arXiv preprint arXiv:2109.02903.

[10] Padmamala, R. ”Word level translation (tamil-english) with word sense disambiguation in tamil using ontnet.” In 2015 International Conference on Computing and Communications Technologies (ICCCT), pp. 191 198. IEEE, 2015.

[11] Kannan, Abishek, Gaurav Mohanty, and Radhika Mamidi. ”Towards building a SentiWordNet for Tamil.” In Proceedings of the 13th Inter national Conference on Natural Language Processing, pp. 30-35. 2016.

[12] J Whalley and N Medagoda. 2015. Sentiment lexicon construction using sentiwordnet 3.0. ICNC’15- FSKD’15, School of Information Science and Engineering, Hunan University, China.

[13] Kannan, Abishek. ”Sentiment lexicon creation in tamil using hybrid techniques.” PhD diss., International Institute of Information Technology Hyderabad, 2019.

[14] Manohar, Kavya. ”Linguistic challenges in Malayalam speech recognition: Analysis and solutions.” PhD diss., College of Engineering Trivandrum, 2023.