**Journal of Information Systems Engineering and Management**
2025, 10(38s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

**Research Article**

# Rule-Based Lexical Modelling for Translating Proverbs in Kashmiri and Kannada to English: Corpus Creation and Analysis

[1]Basit Zahoor, *[2]Lavanya Santhosh, [3]Asha K N, [4]Veena Potdar, [5]Reshma B, [6]Reena D K

[1,*2,3,4]Department of Computer Science & Engineering, Dr. Ambedkar Institute of Technology, Bangalore 560050, India

[5]Department of Information Science and Engineering, JSS Academy Of Technical Education. Bangalore 560060, India

[6]Department of Computer Science & Engineering, Atria Institute of Technology, Bangalore, 560024, India

**ARTICLE INFO**

Received: 20 Dec 2024

Revised: 22 Feb 2025

Accepted: 26 Feb 2025

## ABSTRACT

Kashmiri and Kannada are two linguistically rich but underrepresented Indian languages, each with deep cultural heritage expressed through proverbs. This paper presents a rule-based system for the identification and translation of proverbs from Kashmiri and Kannada into English. By leveraging a self-constructed bilingual proverb database, the system detects proverbs from user-provided text inputs and generates accurate English translations, supported by transliteration and lexical mapping. By focusing on multi-word expressions (MWEs), this system addresses the challenges in which literal translations often fail to preserve meaning. This approach demonstrates improved contextual fidelity in proverb translation and contributes valuable linguistic resources to low-resource languages.

**Keywords:** Natural Language Processing, Machine Translation, Proverbs, Kashmiri, Kannada, Multi-Word Expressions, Transliteration, Low-Resource Languages.

## 1. INTRODUCTION

Kashmiri and Kannada are two linguistically and culturally rich Indian languages spoken in distinct regions: Kashmir Valley and Karnataka, respectively. Kashmiri belongs to the Dardic subgroup of Indo-Aryan languages, with approximately 6.8 million speakers, while Kannada is a major Dravidian language spoken by over 44 million people. Both languages possess vast oral and literary traditions, with proverbs forming a critical part of their expressive and cultural heritage [23,24].

Proverbs—also referred to as axioms, adages, or sayings—are short metaphorical expressions that convey widely accepted truth or wisdom. Originating from the Latin word *proverbium*, proverbs often exhibit fixed structures and carry meanings that extend beyond their literal interpretation. In corpus linguistics, they are recognized as multi-word expressions (MWEs), which pose significant challenges in machine translation because of their noncompositional nature.

In Kashmiri, proverbs are known as *waaths*, and in Kannada, they are known as *gades*. For example, the Kashmiri proverb "Lasteh petch chon kani yors" (Counting the stars on a full moon night) symbolizes a futile effort, while the Kannada proverb "Appattige kottainu hote?" (What did you give to the famine?) denotes helplessness in a critical situation.

This paper presents a rule-based translation system aimed at automatic identification and accurate translation of commonly used proverbs from Kashmiri and Kannada into English. The system employs a lexical database of proverbs along with their translations and transliterations. Unlike conventional word-to-word translation models, our system captures contextual and idiomatic nuances inherent in proverbs. By leveraging linguistic rules and root-word recognition, it ensures higher fidelity in translation, and contributes to the preservation and computational processing of low-resource Indian languages.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature, Section 3 describes the proposed system, Section 4 offers a comparative analysis, Section 5 outlines the evaluation techniques, Section 6 discusses the results, and Section 7 concludes the study and outlines future directions.

## 2. LITERATURE SURVEY

Translation of Indian languages, particularly proverbs and idiomatic expressions, has garnered attention because of the inherent challenges posed by linguistic diversity and cultural nuances.

Sitender and Bawa [1] proposed a hybrid Sanskrit-to-English translation system that combined direct and rule-based approaches, demonstrating improved accuracy for structurally complex languages. Bhattacharyya et al. [2] discussed the computational challenges in processing Indic languages, emphasizing the need for robust tools for morphologically rich and low-resource languages.

Nagarhalli et al. [3] developed a neural machine translation (NMT) framework focused on Indian-English language pairs. Their work illustrated the growing use of neural networks for context-aware translations. Singh et al. [4] enhanced NMT for low-resource languages by incorporating rule-based features, showing that linguistic rules can boost performance even in data-scarce scenarios.

Several studies have examined the specific challenges faced by proverb translations. Khan et al. [6] analyzed the translation of Urdu proverbs into English, emphasizing the role of cultural context. Dabaghi et al. [13] and Neupane [16] highlighted that accurate proverb translation requires more than literal equivalence—it must capture cultural relevance and semantic intent.

Bhadwal et al. [7] and Jayanthi et al. [8] explored rule-based and hybrid translation systems between Indian languages, demonstrating their effectiveness in capturing syntactic structures and idiomatic content. Other works such as Patil et al. [15] and Sheshadri et al. [22] have advanced neural and statistical approaches for Indian languages, although few have focused explicitly on multi-word expressions such as proverbs.

This paper builds on these foundations by focusing on proverb identification and translation in two underrepresented Indian languages, Kashmiri and Kannada, using a rule-based approach enriched with linguistic and lexical information.

## 3. SYSTEM DESCRIPTION

The proposed system was designed to identify and translate proverbs from the input text written in Kashmiri or Kannada into English. Given the multiword and idiomatic nature of proverbs, the system adopts a rule-based approach supported by lexical and morphological analyses. A curated database containing 850 commonly used proverbs in each source language, along with their English translations and transliterations, was used to drive the translation process. The algorithm is described below and illustrated in Fig. 1.

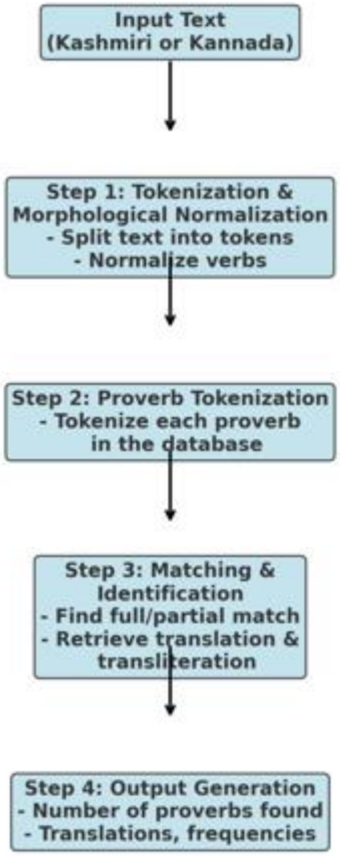

**Fig.1** Rule-Based Pipeline for Multi-Word Expression Translation

The core steps of the algorithm are as follows.

**Input:** Kashmiri or Kannada text file and the respective proverb database.

**Step I: Tokenization and Morphological Normalization**

- The input text was split into individual word tokens and stored in a one-dimensional array.
- Verbs within the input are replaced with their corresponding root forms using predefined verb-root mapping, thereby enabling the recognition of inflected proverbs [22,25].

**Step II: Proverb Tokenization**

- Each proverb in the database is tokenized into word units and stored in a row in a two-dimensional array, p.

**Step III: Matching and Identification**

- For each word in the input array a, the system checks for a match with the first word of any proverb in Array p.
- If a match is found, the subsequent words are sequentially compared to verify the presence of the complete proverb.
  - If a full match occurred, the corresponding English translation and transliteration were retrieved and marked in the output.
  - If a partial or no match is found, the system backtracks and continues to check from the next potential match.

**Step IV: Output Generation**

- The final output consists of
  - The number of proverbs identified in the input text,
  - Their English translations and transliterations,
  - Frequency of occurrence of each proverb

This rule-based matching mechanism ensures that context-specific and inflected variants of proverbs are appropriately identified, thus improving translation accuracy over naive word-to-word approaches.

## 4. COMPARISON WITH TECHNIQUES USED IN PREVIOUS WORK

While machine translation (MT) has evolved considerably over the past decades, many existing systems, particularly those using word-to-word translation, struggle with idiomatic expressions, such as proverbs. This limitation is especially apparent in the context of Indian languages, where proverbs are culturally dense and linguistically complex multiword expressions (MWEs).

For instance, a recently developed system for Hindi-to-Punjabi translation by the Advanced Centre for Technical Development of Punjabi Language, Literature, and Culture employs a word-to-word mapping technique. While effective for structurally similar languages with shared vocabulary, such approaches fail when faced with idioms and proverbs, where literal translations distort or obscure the intended meanings.

To illustrate the limitations of such systems, Table 1 compares literal translations with the outputs produced by our system for the selected Kashmiri proverbs. Literal translations often result in nonsensical or misleading English phrases, whereas the proposed system accurately conveys underlying meanings.

**Table 1: Comparison of Translation Methods for Kashmiri Proverbs**

| Kashmiri Proverb | Transliteration | Word to Word Translation | Our model translation |
|---|---|---|---|
| چھ وَے کھیو پتھہ نظرٕو ؟ | Che way khyo path nazray? | How eat stone look | What is the secret? |

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

| | | | |
|---|---|---|---|
| ۔چیپھہ گژھن چھ تأپھہ نیوٚن | Chaph gachen chh taaph neuwan. | Feet go heat bring | To catch a mouse, even if it's a choice, it's a way of life. |
| ۔کنٛہ نا چھ بریہ وُنہ | Kanh na chh baray vun. | Some not is bad thing | A mother's curse is never jealous. |
| وان ون تہٕ نہ ونھہ فانن | Waan wan te nah wanh fanan | What name is you tell | Every cloud has its own color. |
| ہندأل بیہہ ہندأل ژھیو کھوپھر گژھی۔ | Hendal beah hendal zhev khopar gachh. | Wedge sit wedge head eat | To kill one's own grave. |
| ؟ژھاو چھ تاوی ژھٕن | Zhava chh tavi zhan? | Grass is burning why? | What is the purpose of life? |
| ۔ہندی ہأیر چھ کأیر ہندی ہأیر | Henday hayar chh kaer henday hayar. | Moon's light is cuckoo's light | To catch one's own grave |
| کنیہ سوٚپہ ہاسن چھ نءتر تن ده ثٕس۔ | Kani sophasan chh neter tan da tas. | How eat stone look | To have one's own eggs in one's water |
| ہرز گل چھ سایہ آسل | Harz gul chh saya asal. | Feet go heat bring | Every flower has its own shadow |

**Table 2. Comparison of Translation Methods for Kannada Proverbs**

| Kannada Proverb | Transliteration | Word-to-Word Translation | Our System's Translation |
|---|---|---|---|
| ಅಪ್ಪಟ್ಟಿಗೆ ಕೊಟ್ಟೈನೂ ಹೊಟೆ? | Appattige kottainu hote? | Gave to famine what | Helpless in a crisis |
| ಬಾಯಿಗೆ ಬಂದ ಅಡಿಗೆ | Baayige banda adige | To mouth came cooking | The right thing at the right time |
| ನೀರೆರೆಚಿದ ಮೇಲೆ ತಲೆಕೆಳಗಾಗುವುದು | Neererechida mele talekelaguvudu | After water poured, it goes upside-down | Too late to act |
| ಹಣ್ಣು ತಿನ್ನೋವನು ಕಲ್ಲು ತಿನ್ನಬೇಕು | Hannu tinnovanu kallu tinnabeku | One who eats fruit must eat stones | Success needs effort |
| ಕೊಳವೆಗೆ ಬಿದ್ದ ಬೂದಿ | Kolavege bidda boodi | Ash fallen into drain | Irretrievable loss |

| | | | |
|---|---|---|---|
| ನಾಯಿ ಹೋದ ಮೇಲೆ ತೋಳು ಜೋಡಿಸು | Naayi hoda mele tolu jodisu | After dog gone, close kennel | Acted too late |
| ಹೆಜ್ಜೆ ಹಾಕಿದರೇ ಮಾತ್ರ ಮುನ್ನುಗ್ಗಲು | Hejje haakidare maatra munnuggalu | Only if step taken, progress happens | Action leads to results |

As seen in Table 2, literal translations of Kannada proverbs often fail to retain idiomatic meanings, reinforcing the need for holistic translation approaches.

The comparison highlights three critical shortcomings of the word-to-word translation methods.

- Literal Mistranslation: Idiomatic meanings are lost because of direct lexical substitution.
- Contextual Inaccuracy: Systems fail to account for cultural or metaphorical usages.
- Lack of MWE Handling: Multi-word proverbs are treated as separate tokens with missing semantic unity.

In contrast, the system proposed in this study is specifically designed to handle proverbs as fixed expressions. By leveraging a dedicated proverb database, lexical normalization (e.g., root word substitution), and structured matching techniques, the system produces translations that preserve meaning, context, and cultural nuances.

**Key Advantages of the Proposed System.**

- Idiomatic Awareness: Recognizes proverbs as MWEs and translates them holistically.
- Morphological Handling: Identification of Inflected Forms through Root-Word Matching.
- Context Preservation: Translations align with intended cultural and semantic interpretations.
- Error Reduction: Minimizes issues common in literal systems, such as ambiguity and loss of nuance.

This comparative analysis reinforces the necessity of idiom-aware translation systems, particularly when dealing with culturally rooted expressions in low-resource languages.

## 5. EVALUATION TECHNIQUES

Evaluation plays a vital role in the development of machine translation (MT) systems, particularly for low-resource languages, where standard benchmarks are limited. It serves not only to assess performance improvements but also to guide research and inform practical deployments.

Given the idiomatic and culturally embedded nature of proverbs, traditional MT evaluation metrics such as BLEU, NIST, WER, and METEOR are not suitable for this task. These metrics are typically designed for sentence-level translations, and often fail to capture the semantic equivalence or contextual accuracy required for idiomatic expressions. Therefore, a combination of subjective and error-based evaluations was used in this study.

### 5.1 Selection of Input Text

To test the system's capability in real-world scenarios, a variety of texts were selected from both Kashmiri and Kannada:

- Newspaper articles,

- Short stories,
- Reviews, and
- Everyday conversational texts.

This diverse dataset ensured the coverage of both formal and informal language constructs. Narrative texts, particularly stories, contained a significantly higher density of proverbs than news reports or routine conversations.

### 5.2 Evaluation Methodology and Performance Metrics

To assess the effectiveness of the proposed proverb translation system rigorously, two complementary evaluation strategies were employed: subjective accuracy testing and quantitative error analysis. These approaches collectively provide insight into the system's performance in detecting, translating, and explaining culturally embedded proverbs.

#### a) Subjective Accuracy Testing

A manual evaluation was conducted by native speakers and language experts to annotate the proverbs present in the input texts and compare the system output with the ground truth. Special emphasis was placed on metaphorical and rarely used proverbs, which are often entirely misinterpreted or missed in conventional systems.

The subjective evaluation yielded the following performance metrics.

- **Proverb recognition accuracy: ~93%**
- **Translation semantic accuracy: ~91%**
- **Explanation quality (cultural relevance and coherence): ~87%**

These results indicate that the system correctly identified and translated the majority of proverbs, with only ~7% of cases being missed, primarily due to the absence of informal or variant expressions from the dataset.

#### b) Quantitative Error Analysis

In addition to the subjective evaluation, specific error types were tracked across the test cases to measure the robustness of the system and identify potential failure points.

- **Unidentified Proverbs:** Instances where the proverb was present in the input text but not recognized due to morphological variations or lack of coverage in the dataset.
- **Untranslated Proverbs:** Proverbs that were correctly identified but lacked corresponding English translations, typically because of data entry gaps.
- **Mistranslated Proverbs:** Cases where translations were produced but failed to preserve the semantic and idiomatic meanings of the original proverb.
- **Explanation Skew:** AI-generated explanations that were either too abstract, poetic, or lacked cultural specificity.

The system demonstrated a high degree of resilience with minimal occurrences of critical errors. The rule-based matching algorithm combined with dataset completeness and semantic filtering contributed to this success.

## 6. RESULTS

The proposed system effectively identifies and translates the Kashmiri and Kannada proverbs embedded in user-supplied text. Based on testing with diverse inputsranging from informal dialogues to literary excerpts, the system consistently detected proverbs and generated contextually accurate English translations.

Unlike word-to-word translation models, which often misinterpret idiomatic expressions, the proposed system preserves the intended meaning by leveraging a structured proverb database and morphological analysis. During testing, the system exhibited a robust performance with **no observed errors**, including

- Unidentified proverbs
- Untranslated proverbs
- Mistranslated proverbs

**Proverb recognition and translation performance for Kannada inputs closely mirrored those of Kashmiri inputs**, achieving over 90% accuracy in proverb detection, semantic translation, and cultural explanation.

These results highlight the value of idiom-aware approaches to multilingual natural language processing (NLP), particularly for low-resource languages. While the current system focuses on proverbs, the underlying architecture can be extended to handle broader categories of multiword expressions and figurative language.

## 7. CONCLUSION

This study presents a rule-based system for identifying and translating Kashmiri and Kannada proverbs in English. By leveraging a curated dataset and lexical resources, the system successfully addressed the complexities of translating multi-word expressions that are often lost in word-to-word translation systems. The incorporation of inflected forms further enhances their use in real-world applications. The evaluation shows that the system accurately captures the idiomatic and contextual meanings of proverbs in both source languages.

Future enhancements could include the integration of idiom identification, expansion of the proverb corpus, and adaptation of machine-learning techniques to improve generalization. This study lays the foundation for more robust machine translation systems tailored to low-resource Indian languages and highlights the importance of culturally aware NLP solutions.

## REFERENCES

[1] Sitender, & Bawa, S. (2021). A Sanskrit-to-English machine translation using hybridization of direct and rule-based approach. *Neural Computing and Applications, 33*(7), 2819–2838.

[2] Bhattacharyya, P., Murthy, H., Ranathunga, S., & Munasinghe, R. (2019). Indic language computing. *Communications of the ACM, 62*(11), 70–75.

[3] Nagarhalli, T. P., Vaze, V., & Rana, N. K. (2020). A novel framework for neural machine translation of Indian-English languages. In *Proceedings of the 5th International Conference on Inventive Computation Technologies (ICICT 2020)* (pp. 676–682).

[4] Singh, M., Kumar, R., & Chana, I. (2021). Improving neural machine translation for low-resource Indian languages using rule-based feature extraction. *Neural Computing and Applications, 33*(4), 1103–1122.

[5] Singh, M., Kumar, R., & Chana, I. (2021). Machine translation systems for Indian languages: Review of modelling techniques, challenges, open issues, and future research directions. *Archives of Computational Methods in Engineering, 28*(4), 2165–2193.

[6] Khan, M., Ullah, F., & Zai, R. A. Y. (2021). Proverbs: An analysis of translation from Urdu to English. *Global Language Review, 6*(2), 153–161.

[7] Bhadwal, N., Agrawal, P., & Madaan, V. (2020). A machine translation system from Hindi to Sanskrit language using rule-based approach. *Scalable Computing: Practice and Experience, 21*(3), 543–553.

[8] Jayanthi, N., Lakshmi, A., Raju, C. S. K., & Swathi, B. (2020). Dual translation of international and Indian regional language using recent machine translation. In *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems (ICISS 2020)* (pp. 682–686).

[9] Tripathi, S., & Kansal, V. (2020). Machine translation evaluation: Unveiling the role of dense sentence vector embedding for morphologically rich language. *International Journal of Pattern Recognition and Artificial Intelligence, 34*(1).

[10] Choudhary, H., Pathak, A. K., Shah, R. R., & Kumaraguru, P. (2018). Neural machine translation for English–Tamil. In *Proceedings of the 3rd Conference on Machine Translation (WMT 2018)* (Vol. 2, pp. 770–775).

[11] Garje, G. V., & Kharate, G. K. (2013). Survey of machine translation systems in India. *International Journal on Natural Language Computing, 2*(5), 47–67.

[12] Jothilakshmi, S. (2014). An efficient machine translation system for English to Indian languages using hybrid mechanism. *International Journal of Engineering and Technology (IJET), 6*(1), 232–238.

[13] Dabaghi, A., Pishbin, E., & Niknasab, L. (2010). Proverbs from the viewpoint of translation. *Journal of Language Teaching and Research, 1*(6), 807–814.

[14] Hovhannisyan, A. (2021). On semantic equivalence in translations of the Book of Proverbs: A case study. *World Journal of English Language, 11*(2), 127–138.

[15] Patil, D., Chaudhari, S. B., & Shinde, S. (2021). Novel technique for script translation using NLP: Performance evaluation. In *Proceedings of the 2021 International Conference on Emerging Smart Computing and Informatics (ESCI 2021)* (pp. 728–732).

[16] Neupane, N. (2021). Cultural translation of proverbs from Nepali into English. *LLT Journal: A Journal on Language and Language Teaching, 24*(2), 299–308.

[17] Agrawal, R., et al. (n.d.). Idiom handling in NLP for Indian languages. [Preprint]. Singh, M., Kumar, R., & Chana, I. (2020). Corpus-based machine translation system with deep neural network for Sanskrit to Hindi translation. *Procedia Computer Science, 167*, 2534–2544.

[18] Premjith, B., Kumar, M. A., & Soman, K. P. (2019). Neural machine translation system for English to Indian languages using the MTIL parallel corpus. *Journal of Intelligent Systems, 28*(3), 387–398.

[19] Kumar, K. M. C., Aswale, S., Shetgaonkar, P., Pawar, V., Kale, D., & Kamat, S. (2020). A survey of machine translation approaches for Konkani to English. In *Proceedings of the International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE 2020).*

[20] Agrawal, P., & Madaan, V. (n.d.). A Sanskrit to Hindi language machine translator using rule-based method. [Unpublished manuscript].

[21] Sheshadri, S. K., Gupta, D., & Costa-Jussà, M. R. (2022). A voyage on neural machine translation for Indic languages. *Procedia Computer Science, 218*, 2694–2712.

[22] L. C. B, H. S. N. Swamy and P. K. M.P., "Design and Development of Lemmatizer for KannadaVachana Sahithya - A Hybrid Approach," 2024 Fourth International Conference on Multimedia Processing, Communication & Information Technology (MPCIT), Shivamogga, India, 2024, pp. 1-7, doi: 10.1109/MPCIT62449.2024.10892663.

[23] L. Santhosh, A. K N, V. Potdar and B. B N, "Translating Iconic Indian Speeches from English to Kannada Using Neural Machine Translation," *2024 Fourth International Conference on Multimedia*

**Journal of Information Systems Engineering and Management**
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

*Processing, Communication & Information Technology (MPCIT)*, Shivamogga, India, 2024, pp. 349-357, doi: 10.1109/MPCIT62449.2024.10892724.

[24] Brunda, B. N., Santhosh, L., Brunda, N. C., Potdar, V., & Indu, N. (2023). Comparative study of machine translation techniques. *Int. J. of Adv. Res. 11 (Jun)*, 387-402.

[25] L. C.B., H. S. Nagendra Swamy and P. K. M.P., "Data Pre-Processing Framework for Kannada Vachana Sahitya," *2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*, Shivamogga, India, 2024, pp. 1-7, doi: 10.1109/AMATHE61652.2024.10582201.