A Word-Splitting Approach to Sanskrit Sandhi Words of Kannada Useful in Effective English Translation

Anonymous ACL submission

Abstract

Natural Language Processing is a field of artificial intelligence that facilitates man-machine interactions through vernacular languages. There are two types of Sandhi in the Kannada Language: Kannada Sandhi and Sanskrit Sandhi. A morph-phonemic word 'Sandhi' is formed when two words or distinct morphemes are joined or combined. A Sandhi word splitting is the reverse of the process of formation. The rules govern Sandhi words in all the Dravidian languages. A rule-based splitting method is developed to obtain the constituent words from the Sanskrit Sandhi words in Kannada sentences. Once the Sanskrit Sandhi (SS) words are split, the type of Sandhi is also identified, leading to an effective translation of the Sanskrit Sandhi words into English. This paper covers seven types of SS words: Savar-Nadeergha, YaN, GuNa, Vruddhi, Jatva, Shchutva and Anunasika Sandhi. The identified split points are as per the Sandhi rules. A dataset of 4900 Sanskrit Sandhi words occurring in Kannada sentences is used to assess the performance of the proposed method, which has given an accuracy of 90.03% and 85.87% in Sanskrit Sandhi identification and in an acceptable English translation. The work finds applications in other Dravidian languages.

1 Introduction

Natural Language Processing (NLP) makes computers understand any language humans speak in the real world, such as English, Hindi, Marathi, Tamil, Telugu, Kannada, Punjabi, etc. NLP helps machines comprehend human interactions. This involves separating words from sentences as per the word boundaries (Vempaty and Nagalla, 2011). Language establishes communication for humans. The language grammar gives language structure and is a system of rules that governs a language's correctness and compliance (Caryappa et al., 2020). Dravidian languages are a family of around 70 languages spoken by nearly 200 million people in different parts of India and the world. Tamil, Malayalam, Kannada and Telugu, and over 20 non-literary languages, are standard in India (Krishnamurthy, 2024). Kannada is one of the major Dravidian languages of India, spoken predominantly in the state of Karnataka, with a 2500-year-old rich cultural history (Amarappa and Sathyanarayana, 2015). It is the world's 27th most widely spoken language, with about 35 million speakers. It has poor resources and considerable syntactic and semantic variance. Kannada is not explored much in Machine Translation (MT) compared to other Indian languages (Nagaraj et al., 2021) and offers more challenges. Table 1 gives the number of speakers of Dravidian languages state-wise and worldwide.

Kannada has a linguistic construct called Sandhi (संधि in Sanskrit, ಸಂಧಿ in Kannada) wherein two words or morphemes merge, causing phonetic or morphological changes at the word's junction. This transformation is seen in many Indian languages, including Sanskrit, Telugu, Tamil, etc., and is governed by the specific grammatical rules. The word Sandhi is used in singular and plural forms throughout this paper. Splitting is the process of obtaining the constituent words from Sandhi word and converting the Sandhi word to an equivalent English (Natarajan and Charniak, 2011). Sandhi splitting approaches are broadly categorized into Dictionary-based, Rule-based, and Corpus-based (Shashirekha and Vanishree, 2016). There are two types of Sandhi in the Kannada Language: Kannada Sandhi and Sanskrit Sandhi. There

Language	Speakers	Locations
Telugu	83,000,000	Andhra Pradesh, Telangana and parts of Karnataka, UK, USA, Australia, Canada, UAE
Tamil	77,000,000	Tamil Nadu, Parts of Karnataka, Maharashtra, Kerala, France, Germany, Italy, USA, UK
Kannada	45,000,000	Karnataka, Kerala, Tamil Nadu, Maharashtra, USA, UK, Canada, UAE, Saudi Arabia
Malayalam	37,000,000	Kerala, Tamil Nadu, Maharashtra, Karnataka
Tulu	1,850,000	Karnataka, Kerala, Gujarat, Saudi Arabia
Beary	1,500,000	Karnataka, Kerala, Gulf Countries
Brahui	2,430,000	Baluchistan (Pakistan), Helmand (Afghanistan)

Table 1: Speakers of Dravidian Languages

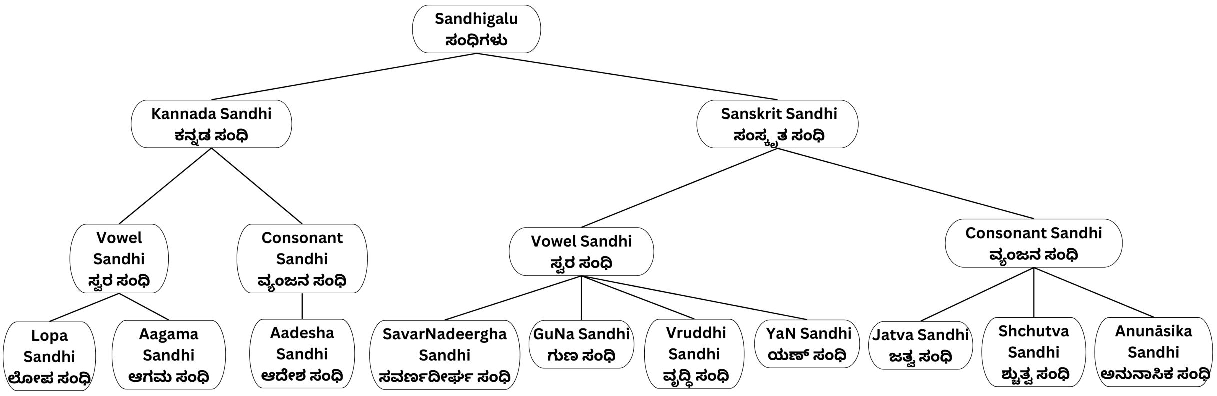


Figure 1: Classification of Sandhi Forms.

are three types of Kannada Sandhi: Lopa Sandhi, Aagama Sandhi, and Aadesh Sandhi. In Sanskrit Sandhi, there are seven types such as SavarNadeergha Sandhi, GuNa Sandhi, YaN Sandhi, Vruddhi Sandhi, Jatva Sandhi, Shchutva Sandhi, and Anunasika Sandhi. The classification of Sandhi forms in the Kannada language is shown in Figure 1. This paper presents a work on the Sanskrit Sandhi helpful in translating Kannada text into English, as part of the contribution to Machine Translation (MT).

MT bridges language barriers and is considered challenging for languages with complex linguistic structures like Kannada. The challenges in MT are related to grammar, while others are related to language generation, multilingual dictionaries, word analysis, etc. (Alawneh and Sembok, 2011). Some of the existing translators, like Google, Bing, Quillbot, i-Translate, etc., do not give satisfactory translations of sentences with Sandhi words. For example, the Kannada sentence "ಯೋಗ ಮತ್ತು ಧ್ಯಾನ ಮನಶ್ಚಂಚಲತೆಯನ್ನು ಕಡಿಮೆ ಮಾಡುತ್ತದೆ" and its transliteration (TL) is "Yoga mattu dhyana manaschanchalatheyannu kaDime maDuttade". Its English translation should be 'Yoga and meditation reduce the boggling mind'. But when we subjected this sentence to the existing translators, which failed to translate the given Kannada sentence, having the Sanskrit Shchutva Sandhi word "ಮನಶ್ಚಂಚಲತೆ", its transliteration (TL) form 'manaschanchalate'. Hence, the present paper provides a devised rule-based Sandhi splitting method useful for converting Sanskrit Sandhi words to English, thereby effectively translating Kannada sentences to English.

2 Literature Survey

A literature survey was conducted to learn about state-of-the-art Sandhi splitting, identification, and machine translation methods.

Information retrieval (IR) in languages with complex morphological patterns, Indian languages, requires breaking down compound words (also called de-compounding) into their parts. The corpus-based models were used extensively for de-compounding, requiring subtle assistance of semantics and sparsity (Sahu and Pal, 2024). The machine learning models were implemented using recurrent neural networks,

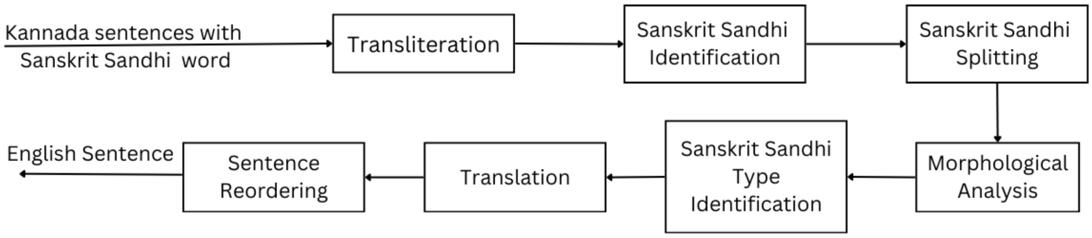
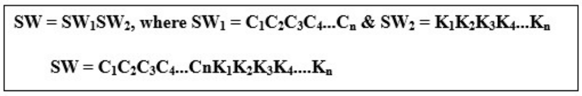


Figure 2: Block Schematic Diagram of Proposed Methodology

long-short-term memory models, and double decoder models. (S. et al., 2024). The morphological analysis of Sanskrit Sandhi words was context-dependent, and Sandhi split, also known as "Vichchhed", was a challenging task. The existing methods include the predetermined splitting rules. However, finding the exact split point is important and determines accuracy issues (Phadke and Patankar, 2023). Nine methods were deployed for the "Sandhi" Splitter: the Bayesian Word Segmentation method, Conditional Random Field, Recurrent Neural Network, Hidden Markov Model, Rule-Based Approach (RBA), Deep Learning, Machine Learning, and Finite State Automata. Sandhi splitters were developed by researchers using RBA (Gaikwad and Saini, 2021). Recurrent Neural Networks (RNNs) were widely used to perform machine translation. A mix of Naïve Bayes and LSI (Latent Semantic Indexing) predicted the next word in Kannada translation. The model was trained using a variety of patterns created by combining bigram, trigram, and 4-gram to improve accuracy (Nandini et al., 2020). The problem was a sequence-to-sequence prediction task and used modern deep-learning techniques. A compound-word (Sandhi) generation and splitting in the Sanskrit Language using LSTM and Bi-LSTM techniques was carried out, and a good prediction accuracy was achieved (Dave et al., 2020). The use of data and grammatical rules of Sanskrit played a significant role in splitting Upasarga and Pratyaya (Angle et al., 2018). The end-to-end neural network models resolved phonetic merges to tokenize Sanskrit (Sandhi) words. The character-level recurrent and convolutional neural networks helped segment words in Sanskrit (Hellwig and Nehrdich, 2018). The literature survey reveals that few authors have worked on Sandhi splitting for languages such as Telugu, Sanskrit, Malayalam, and Kannada. Not all Sandhi are considered, and works have emphasised one or two types of Kannada Sandhi. Sanskrit Sandhi is not explored much. Hence, the present paper deals with an account of the translation of the Sanskrit Sandhi words of the Kannada language to English. It is a complete work encompassing all kinds of SS and the rules for splitting into constituent words.



SW = SW_1SW_2, where $SW_1 = C_1C_2C_3C_4...C_n$ & $SW_2 = K_1K_2K_3K_4...K_n$

SW = $C_1C_2C_3C_4...CnK_1K_2K_3K_4....K_n$

Box 1: Structure of Sanskrit Sandhi Word

3 Dataset Preparation and Proposed Methodology

The required dataset is collected and prepared for testing the method. The block schematic diagram of the stages of processing is discussed.

3.1 Dataset Preparation

The data is collected from some Kannada storybooks and input from native Kannada language speakers Kuvempu (1971) and Keshiraja (1920). The dataset comprises 4900 Sanskrit Sandhi words drawn from Kannada sentences containing one word, two words, and three words of Sanskrit Sandhi, as given in Table 2.

3.2 Proposed Methodology

The proposed methodology is divided into seven phases: Transliteration, Sanskrit Sandhi identification, Sanskrit Sandhi Word Splitting, Morphological Analysis, Sanskrit Sandhi Type Identification, Translation and Sentence Reordering, as shown in Figure 2.

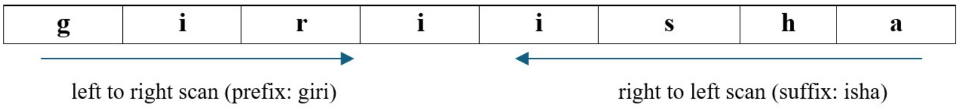

g
i
r
i
i
s
h
a
left to right scan (prefix: giri)
right to left scan (suffix: isha)

Figure 3: Prefix-Suffix method

Sentences	Count
Total No. of Sanskrit Sandhi words	4900
Total No. of Kannada Sentences	3736
No. of sentences without Sandhi words	39
Sentences having one Sandhi word	2768
Sentences having two Sandhi words	655
Sentences having three Sandhi words	274

Table 2: Sanskrit Sandhi Dataset

3.2.1 Transliteration

Transliteration (TL): It is a phonetic resemblance way of writing, converting words from one language script to another by putting them in a familiar alphabet. Romanization transliterates the vowels(KV) and consonants(KC) of Kannada, as given in Tables 3 and 4, respectively. Transliteration changes the characters from the word's original alphabet to similar-sounding characters in a different script.

3.2.2 Sanskrit Sandhi Identification

The sentences are subjected to tokenization. The tokens obtained are checked against a dictionary of root words to determine whether the token is a Sandhi word. The Sandhi words are identified based on their transformations. Let SW be the given Sanskrit Sandhi word, which is the concatenation of two words, namely SW_1 and SW_2, represented as SW=SW_1SW_2 where SW_1 and SW_2 are the the two constituent words with sequences of characters as defined by expressions (1) and (2).

$$SW_1 = C_1C_2C_3C_4 \cdots C_n \qquad (1)$$

$$SW_2 = K_1K_2K_3K_4 \cdots K_n \qquad (2)$$

Let C_i and K_i represent the i$^{\text{th}}$ character in words SW_1 and SW_2, respectively, and i = 1,2,3....n describe the characters in the words SW_1 and SW_2. The word SW can be written as shown in Box 1.

3.2.3 Sanskrit Sandhi Word Splitting

Sandhi Word splitting (SWS), also called Sandhi Vichchheda, is a technique to split a string of conjoined words into a sequence of constituent root words. We have maintained the dictionaries of prefixes, suffixes, and root words in DWAG (Directed word acyclic graph) structure. We have used the prefix-suffix method for Sandhi Word Splitting. In the proposed prefix-suffix Sandhi Word Splitting method, the Sandhi word undergoes character-by-character scanning, in both directions, resulting in prefix and suffix words. The SWS involves scanning in two directions: left-to-right to identify the prefix word, which is further verified against a corresponding dictionary, and right-to-left to determine the suffix word, which is subsequently validated using the suffix dictionary, as shown in Figure 3. For example, the split of the word "ಗಿರೀಶ" (TL: giriisha) is shown in Figure 3. The given word will be split as ಗಿರೀಶ (TL: giriisha) => ಗಿರಿ (TL: giri) + ಈಶ (TL: isha) by scanning from left to right and right to left, respectively.

3.2.4 Morphological Analysis

Morphological analysis is used to identify all the morphemes from agglutinative words and their grammatical categories. This helps to improve the understanding of a language's word structure and meaning. Morphological analysis helps accurately identify and reconstruct the original Sandhi words. Morphological analysis is crucial to Machine Translation (MT) and improves the translation accuracy, especially for morphologically rich languages like Kannada. Since Kannada words often contain complex prefixes, suffixes, and Sandhi combinations, breaking them down correctly helps in meaningful translation into English or other languages.

3.2.5 Sanskrit Sandhi Type Identification

The Sandhi words are split, and the rules are applied to find the category of a Sandhi. The

KV	TL	KV	TL	KV	TL	KV	TL	KV	TL	KV	TL
ಅ	a	ಆ	aa,A	ಇ	i	ಈ	ee, I, ii	ಊ	u	ಊ	oo, U, uu
ಋ	Ru	ಎ	e	ಏ	ai, ei	ಒ	o	ಓ	O	ಔ	au, ou

Table 3: Romanization of Kannada Vowels

KC	TL	KC	TL	KC	TL	KC	TL	KC	TL
ಕ	ka, qa	ಚ	ca, cha	ಟ	Ta	ತ	ta	ಪ	pa, fa, pha
ಖ	Ka, kha	ಛ	Ca	ಠ	Tha	ಥ	tha	ಫ	Pa
ಗ	ga	ಜ	ja	ಡ	Da	ದ	da	ಬ	ba
ಘ	Ga	ಝ	Ja, jha	ಢ	Dha	ಧ	dha	ಭ	Ba, bha
ಙ	ga	ಞ	ja	ಣ	Na	ನ	na	ಮ	ma
ಯ	ya	ರ	ra	ಲ	la	ಳ	La	ವ	va, wa
ಶ	Sa	ಷ	Sha	ಸ	sa	ಹ	ha		

Table 4: Romanization of Kannada Consonants

Sandhi word is valid if it can be split into a prefix and a suffix. It is possible to identify the Sandhi split point by applying Kannada grammar rules, and the category of Sandhi (Aralikatte et al., 2018); (Gopal Krishna Udupa N, 2020); (Keshiraja, 1920).

i. Sanskrit Sandhi Rules There are seven types of Sanskrit Sandhi in Kannada and each Sandhi is governed by definite rule for joining the two constituent words. Following are the rules devised for the Sanskrit Sandhi.

Rules	Split Words	Sandhi Word
a + a -> aa ಅ + ಅ ->ಆ	deva + asura ದೇವ + ಅಸುರ	devaasura ದೇವಾಸುರ
aa + a -> aa ಆ + ಅ -> ಆ	vidyaa + abhyasa ವಿದ್ಯಾ + ಅಭ್ಯಾಸ	ivdyaabyasa ವಿದ್ಯಾಭ್ಯಾಸ
i + i -> i ಇ + ಇ -> ಈ	kavi + iMdra ಕವಿ + ಇಂದ್ರ	kaviiMdra ಕವೀಂದ್ರ
u + u -> uu ಉ + ಉ -> ಊ	vadhu + upadesha ವಧು + ಉಪದೇಶ	vadhuupadesha ವಧೂಪದೇಶ
i + ii -> ii ಇ + ಈ ->ಈ	giri + iisha ಗಿರಿ + ಈಶ	giriisha ಗಿರೀಶ

Table 5: SavarNadeergha Sandhi Rules with Examples

- **SavarNadeergha Sandhi**: When two vowels occur in a word, one after the other, a single long vowel is substituted for both. This is called an extended vowel conjugation. The rules with sample examples are given in Table 5.

- **Vruddhi Sandhi**: If the prefix ends with characters 'a', and 'aa', and the suffix begins with characters 'i', 'ai', or 'au', during the sandhi word formation, these are replaced by 'ai' and 'au', respectively. The rules with sample examples are given in Table 6.

Rules	Split Words	Sandhi Word
a + i -> ai ಅ + ಏ -> ಐ	loka + ikya ಲೋಕ + ಏಕ್ಯ	lokaikya ಲೋಕೈಕ್ಯ
aa + ai -> ai ಆ + ಐ -> ಐ	vidyaa + aishwarya ವಿದ್ಯಾ + ಐಶ್ವರ್ಯ	vidyaishwarya ವಿದ್ಯೈಶ್ವರ್ಯ
a + au -> au ಅ + ಔ ->ಔ	Ghana + audharya ಘನ + ಔಧಾರ್ಯ	Ghanaudharya ಘನೌದಾರ್ಯ
aa + au -> au ಆ + ಔ ->ಔ	mahaa + audharya ಮಹಾ + ಔಧಾರ್ಯ	mahaudhrya ಮಹೌಧಾರ್ಯ

Table 6: Vruddhi Sandhi Rules with Examples

- **GuNa Sandhi**: If the prefix ends with characters 'a' and 'aa' and the suffix begins with characters 'i', 'u', and 'ru', then the letters 'e', 'oo', and 'r' will be replaced in the sandhi formation. This is called 'GuNa' Sandhi. The rules with sample examples are given in Table 7.

- **Jatva Sandhi**: The consonants 'k', 'ch', 'T', 'th', 'p' at the end of the prefix word are replaced by the third consonants of the same class ('g', 'j', 'D', 'd', 'b'). This is called 'Jatva' Sandhi The resulting Sandhi word and the governing rules with sample examples are given in Table 8.

- **YaN Sandhi**: When a sandhi is formed and if the prefix ends with characters 'i', 'u', and 'ru', then the character 'y' replaces ’i’, the character ’v’ replaces ’u’, and the character ’r’ replaces the character ’ru’. This is called a ’YaN’ Sandhi. The rules with sample examples are given in Table 9.

- **Anunasika Sandhi**: The consonants ’k’, ’t’, ’T’, and ’p’ at the end of the prefix word will be replaced with ’gm’, ’na’, ’Na’, and ’ma’ in sandhi formation. The rules with sample examples are given in Table 10.
- **Shchutva Sandhi**: The prefix word has ’s’ or ’th’ as ending characters, and the suffix word has ’sha’ and ’cha’ as beginning characters; then these are replaced by ’sha’, or ’shcha’ and ’chh’ in sandhi formation, respectively. This is called the ’Shchutva’ Sandhi. The rules with sample examples are given in Table 11.

Rules	Split Words	Sandhi Word
a + i -> e ಅ +ಇ -> ಏ	sura + iMdra ಸುರ + ಇಂದ್ರ	sureNdra ಸುರೇಂದ್ರ
aa + i -> e ಆ + ಇ -> ಏ	dharaa + iMdra ಧರಾ + ಇಂದ್ರ	dhareNdra ಧರೇಂದ್ರ
a + u -> oo ಅ + ಉ ->ಊ	soorya + udaya ಸೂರ್ಯ + ಉದಯ	sooryoodaya ಸೂರ್ಯೋದಯ
a + ru -> ar ಅ + ಋ ->ರ್	deva + rushi ದೇವ + ಋಷಿ	devarshi ದೇವರ್ಷಿ
aa + ru -> ar ಆ + ಋ -> ರ್	mahaa + rushi ಮಹಾ + ಋಷಿ	maharshi ಮಹರ್ಷಿ

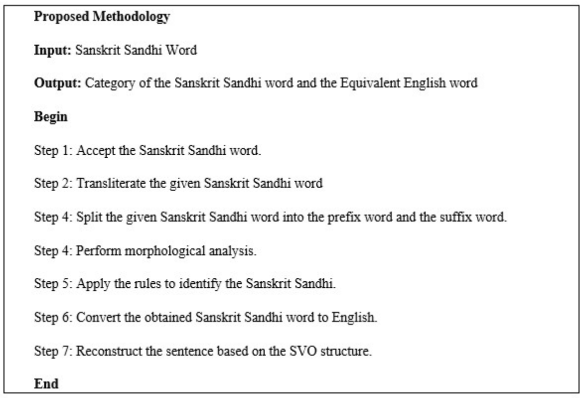
Table 7: GuNa Sandhi Rules with Examples

Rules	Split Words	Sandhi Word
k -> g ಕ -> ಗ	vak + iisha ವಾಕ್ + ಈಶ	vageesha ವಾಗೀಶ
ch -> j ಚ ->ಜ	ach + aadi ಅಚ್ + ಆದಿ	ajaadi ಆಜಾದಿ
T ->D ಟ -> ಡ	viraaT + roopa ವಿರಾಟ್ + ರೂಪ	viraaDroopa ವಿರಾಡ್ರೂಪ
t-> d ತ -> ದ	sat + uddesha ಸತ್ + ಉದ್ದೇಶ	saduddesha ಸದುದ್ದೇಶ

Table 8: Jatva Sandhi Rules with Examples

Rules	Split Words	Sandhi Word
i + a -> ya ಇ + ಅ ->ಯ	ati + avasara ಅತಿ + ಅವಸರ	atyavasara ಅತ್ಯವಸರ
i + aa -> yaa ಇ + ಆ -> ಯಾ	jaati + aatita ಜಾತಿ + ಆತೀತ	jaatyaatita ಜಾತ್ಯಾತೀತ
i + u -> yu ಇ + ಉ -> ಯು	prati + uttara ಪ್ರತಿ + ಉತ್ತರ	pratyuttara ಪ್ರತ್ಯುತ್ತರ
u + a -> va ಉ + ಅ -> ವ	manu + aadi ಮನು + ಆದಿ	manvaadi ಮನ್ವಾದಿ
ru + a -> ra ಋ + ಅ -> ರ	pitru + aajne ಪಿತೃ + ಆಜ್ಞೆ	pitraajne ಪಿತ್ರಾಜ್ಞೆ

Table 9: YaN Sandhi Rules with Examples



```
Proposed Methodology

Input: Sanskrit Sandhi Word

Output: Category of the Sanskrit Sandhi word and the Equivalent English word

Begin

Step 1: Accept the Sanskrit Sandhi word.

Step 2: Transliterate the given Sanskrit Sandhi word

Step 4: Split the given Sanskrit Sandhi word into the prefix word and the suffix word.

Step 4: Perform morphological analysis.

Step 5: Apply the rules to identify the Sanskrit Sandhi.

Step 6: Convert the obtained Sanskrit Sandhi word to English.

Step 7: Reconstruct the sentence based on the SVO structure.

End
```

Box 2: Overall Proposed Methodology

Rules	Split Words	Sandhi Word
k -> gm ಕ್ -> ಙ್ಮ	vaak + maya ವಾಕ್ + ಮಯ	vaagmaya ವಾಙ್ಮಯ
t -> n ತ್ --> ನ	cit + maya ಚಿತ್ + ಮಯ	chinmaya ಚಿನ್ಮಯ
T -> N ಟ್ --> ಣ	shaT + maasa ಷಟ್ + ಮಾಸ	shaNmaasa ಷಣ್ಮಾಸ
p -> m ಪ್ --> ಮ	ap + maya ಅಪ್ + ಮಯ	ammaya ಅಮ್ಮಯ

Table 10: Anunasika Sandhi Rules with Examples

3.2.6 Translation and Sentence Reordering

In machine translation (MT), four methods, namely Hybrid, Rule-Based, Neural, and Statistical, are deployed, and it is true for Kannada to its equivalent English. We have developed a rule-based machine translation method in the proposed approach that uses specialised dictionaries and Kannada grammar. For example the sentence "ಅವನು ಗಿರೀಶ ಇರುತ್ತಾನೆ" (TL: avanu giriisha iruttane). In this example, the word ಗಿರೀಶ (TL:giriish) is extracted and split. The prefix ಗಿರಿ (TL:giri) and suffix ಈಶ (TL: isha) are obtained using the sandhi splitting method. The meaning of ’giri’ means "mountain" and the meaning of ’isha’ means

Rules	Split Words	Sandhi Word
s + sha -> sha ಸ್ +ಶ ->ಶ	payas + shayana ಪಯಸ್ + ಶಯನ	payashayana ಪಯಶಯನ
s + cha -> shcha ಸ್ + ಚ -> ಶ್ಚ	manas + chanchala ಮನಸ್ + ಚಂಚಲ	manashchanchala ಮನಶ್ಚಂಚಲ
th + cha -> chh ತ್ + ಚ -> ಚ್ಚ	sharath + chaMdra ಶರತ್ + ಚಂದ್ರ	sharachhaMdra ಶರಚ್ಚಂದ್ರ

Table 11: Shchutva Sandhi Rules with Examples

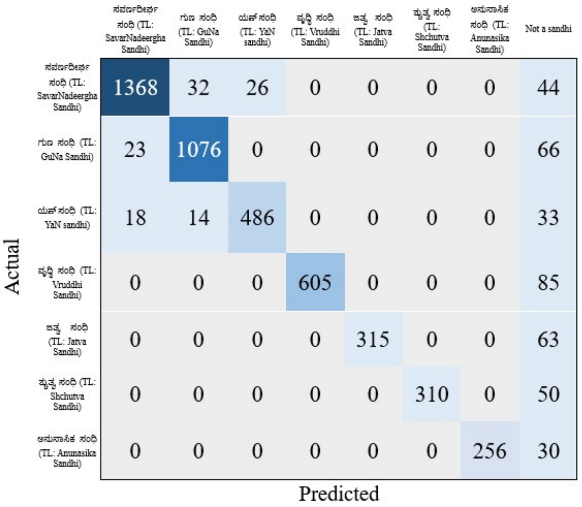


Figure 4: Confusion Matrix for Sanskrit Sandhi Splitting and Identification

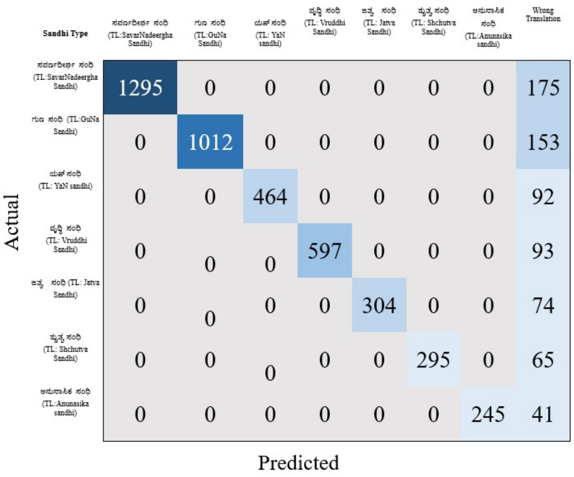


Figure 5: Confusion Matrix for Sanskrit Sandhi Translation

"lord". It is the name of lord Shiva in Hinduism.

In sentence reordering, each non-Sandhi words meaning is obtained in English using the INLTK (Indic Natural Language Toolkit), whereas the Sandhi words need splitting for correct translation. The words in the English sentence are tagged with PoS and reordered according to the SVO structure. Hence, we obtain the effective English translation as "He is the mountain lord", the lord Shiva. The overall Machine translation methodology is given in Box 2.

4 Results of the Proposed Methodology

The proposed method is tested on a corpus of 3736 Kannada sentences containing 4900 Sandhi words, and the performance parameters are computed. The methodology is implemented in Python using the INLTK. The method's accuracy(SIT) is defined as the average percentage of Sandhi words correctly identified (SI) and the percentage of Sandhi words correctly translated into English (ST) as given in expression 3.

$$\%SIT = \frac{\%SI + \%ST}{2} \qquad (3)$$

A confusion matrix (CM) is obtained to determine how well the developed methodology compares with the desired or Actual outcomes. The CM for Sandhi identification and translation is shown in Figures 4 and 5. We have obtatined 90.03% (SI), 85.87% (ST), and 87.95% (SIT) for Sanskrit Sandhi identification and translation. The performance parameters such as Precision, Recall, F1-score and accuracy obtained are given in Tables 12 and 13. The Sanskrit Sandhi Identification and Translation results are shown in Figures 6 and 7 respectively.

5 Conclusion

The developed Rule-Based Methodology (RBM) for Sanskrit Sandhi splitting, identification, and English translation is tested on a corpus of 3736 Kannada sentences containing

Class Name	Precision	Recall	F1-score	Accuracy
ಸವರ್ಣದೀರ್ಘ ಸಂಧಿ (TL: SavarNadeergha Sandhi)|	0.93	0.97	0.95	0.95
ಗುಣ ಸಂಧಿ (TL: GuNa Sandhi)|	0.92	0.96	0.94	0.92
ಯಣ್ ಸಂಧಿ (TL: YaN Sandhi)	0.88	0.95	0.91	0.90
ವೃದ್ಧಿ ಸಂಧಿ (TL: Vruddhi Sandhi)	0.91	1	0.95	0.91
ಜಶ್ತ್ವ ಸಂಧಿ (TL:Jatva Sandhi)	0.83	1	0.91	0.84
ಶ್ಚುತ್ವ ಸಂಧಿ (TL: Shchutva Sandhi)|	0.89	1	0.93	0.86
ಅನುನಾಸಿಕ ಸಂಧಿ (TL: Anunasika Sandhi)|	1	1	1	0.90
Overall|	0.91	0.98	0.94	0.90

Table 12: Sanskrit Sandhi Identification Performance Parameters

Class Name	Precision	Recall	F1-score	Accuracy
ಸವರ್ಣದೀರ್ಘ ಸಂಧಿ (TL: SavarNadeergha Sandhi)	0.88	0.98	0.93	0.88
ಗುಣ ಸಂಧಿ (TL: GuNa Sandhi)	0.85	1	0.92	0.87
ಯಣ್ ಸಂಧಿ (TL: YaN Sandhi)	0.83	1	0.91	0.84
ವೃದ್ಧಿ ಸಂಧಿ (TL: Vruddhi Sandhi)	0.87	1	0.93	0.87
ಜತ್ವ ಸಂಧಿ (TL:Jatva Sandhi)	0.80	1	0.89	0.81
ಶ್ಚುತ್ವ ಸಂಧಿ (TL: Shchutva Sandhi)	0.82	1	0.90	0.82
ಅನುನಾಸಿಕ ಸಂಧಿ (TL: Anunasika Sandhi)	0.86	1	0.92	0.86
Overall|	0.84	0.99	0.91	0.85

Table 13: Sanskrit Sandhi Translation Performance Parameters

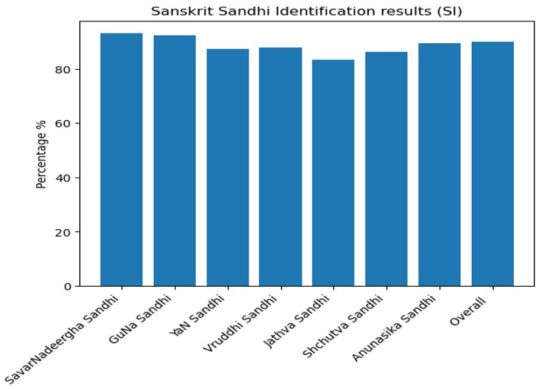


Figure 6: Sanskrit Sandhi Identification(SI) Results

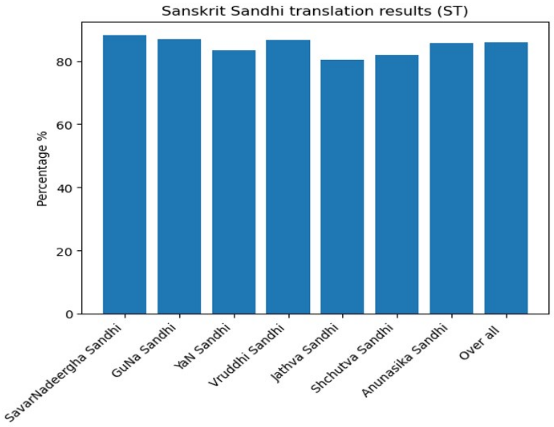


Figure 7: Sanskrit Sandhi Translation(ST) Results

4900 Sanskrit Sandhi words. It has given satisfactory results for the Sanskrit Sandhi such as SavarNadeergha Sandhi, GuNa Sandhi, YaN Sandhi, Vruddhi Sandhi, Jatva Sandhi, Shchutva Sandhi and Anunasika Sandhi. RBM has given an average accuracy of 90.03% for effective identification and 85.87% for translating Sanskrit Sandhi words to English. It is observed that the accuracy of the RBM could be increased with the enhanced dataset and the corresponding prefix and suix words dictionaries. INLTK Toolkit is used for implementation of the proposed methodology. There is a scope to use statistical and deep learning-based methods, and the authors wish to try them in future work. This methodology is helpful for Sandhi splitting in other Dravidian languages.

Limitations

The work presented in this paper is limited to all types of Sanskrit Sandhi and Sanskrit Sandhi words present in Kannada sentences and their effective translations. With an increase in the dataset and the dictionary size, the performance of the proposed methodology could be enhanced.

Ethics statement

This work presents a rule-based method for splitting Sanskrit Sandhi words in Kannada to support effective English translation. While the approach is based on linguistic rules and demonstrates high accuracy, it may not fully capture context-sensitive or culturally significant expressions. Care should be taken when applying the system to religious or literary texts. The dataset used in this work was created by the authors and consists solely of Kannada words and synthetically generated example sentences. We have carefully ensured that it contains no personally identifiable information or offensive content.

References

Mouiad Fadiel Alawneh and Tengku Mohd Sembok. 2011. Rule-based and example-based machine translation from english to arabic. In *Proceedings - 2011 6th International Conference on Bio-Inspired Computing*, pages 343–47. Doi:10.1109/BIC-TA.2011.76.

S. Amarappa and S. V. Sathyanarayana. 2015. Kannada named entity recognition and classification (nerc) based on multinomial naïve Bayes (mnb) classifier. *International Journal on Natural Language Computing*, 4(4):39–52. Doi:10.5121/ijnlc.2015.4404.

Sachi Angle, B. Ashwath Rao, and S. N. Muralikrishna. 2018. Kannada morpheme segmentation using machine learning. *International Journal of Engineering and Technology(UAE)*, 7(2):45–49. Doi:10.14419/IJET.V7I2.31.13395.

Rahul Aralikatte, Neelamadhav Gantayat, Naveen Panwar, Anush Sankaran, and Senthil Mani. 2018. Sanskrit sandhi splitting using seq2(seq)22. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. Vol. 2*, pages 4909–14.

B. C. Caryappa, Vishwanath R. Hulipalled, and J. B. Simha. 2020. Kannada grammar checker using lstm neural network. In *Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020*, pages 332–37. Doi:10.1109/ICSTCEE49637.2020.9277479.

Sushant Dave, Arun Kumar Singh, A. P. Prathosh, and Brejesh Lall. 2020. Neural compound-word (sandhi) generation and splitting in sanskrit language.

Hema Gaikwad and Jatinderkumar R. Saini. 2021. On state-of-the-art of pos tagger, 'sandhi' splitter, 'alankaar' finder and 'samaas' finder for indoaryan and dravidian languages. *International Journal of Advanced Computer Science and Applications*, 12(4):429–36. Doi:10.14569/IJACSA.2021.0120455.

Gopal Krishna Udupa N. 2020. *Kannada Vyakarana Mattu Rachane*. Mcc Publications, 2016th ed.

Oliver Hellwig and Sebastian Nehrdich. 2018. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2754–63. Doi:10.18653/v1/d18-1295.

Keshiraja. 1920. *Shabdamani Darpanam*. Karnataka Sahitya Parishat, Bengalore.

Bhadriraju Krishnamurthy. 2024. Dravidian languages | history, grammar, map, facts | britannica. *Retrieved*, 2024. Https://www.britannica.com/topic/Dravidian-languages.

Pushpalatha Kadavigere Nagaraj, Kshamitha Shobha Ravikumar, Mydugolam Sreenivas Kasyap, Medhini Hulumakki Srinivas Murthy, and Jithin Paul. 2021. Kannada to english machine translation using deep neural network. *Ingenierie Des Systemes d'Information*, 26(1):123–27. Doi:10.18280/isi.260113.

B. R. Nandini, M. Prof Hamsaveni, and V. Prof Charunayana. 2020. Hybrid machine learning based kannada next word prediction. *International Research Journal of Engineering and Technology (IRJET)*, 7:5605–8.

Abhiram Natarajan and Eugene Charniak. 2011. S3 - statistical samdhi splitting. In *IJCNLP 2011 - Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 301–8.

Madhura Phadke and Shreya Patankar. 2023. Exploring the intricacies of sandhi in sanskrit: Phonological rules and linguistic significance. *International Journal of Applied Engineering Technology*, 5(1):353–60.

H. S., Alok Nath M. Sreedeepa, Ajay K. Mani, C. Arun Kumar, and Sumam Mary Idicula. 2024. Review on sanskrit sandhi splitting using deep learning techniques. *Journal of Information Technology and Digital World*, 6(2):136–52. Doi:10.36548/jitdw.2024.2.003.

Siba Sankar Sahu and Sukomal Pal. 2024. A case study on decompounding in Indian language ir. *Natural Language Processing*, pages 1–31. Doi:10.1017/nlp.2024.16.

Vidwan N. Ranganath Sharma. 2010. *Vyakarana-Hosagannada*, 1 edition. Kannada Sahitya Parishat.

H. L. Shashirekha and K. S. Vanishree. 2016. Rule based kannada agama sandhi splitter. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 549–53. IEEE.

Phani Chaitanya Vempaty and Satish Chandra Prasad Nagalla. 2011. Automatic sandhi spliting method for telugu, an Indian language. *Procedia - Social and Behavioral Sciences*, 27(Pacling):218–25. Doi:10.1016/j.sbspro.2011.10.601.