PeerJ Computer Science

# Comparative analysis of machine translation for Hindi-Dogri text using rule-based, statistical, and neural approaches

Joginder Kumar[1], Manik Rakhra[2], Preeti Dubey[3], Deepak Prashar[2,4], Leo Mrsic[5], Arfat Ahmad Khan[6], Seifedine Kadry[7,8] and Jungeun Kim[9]

[1] School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India
[2] Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India
[3] Department of Computer Science and Engineering, Government College for Women, Parade, Jammu, India
[4] Jadara University Research Center, Jadara University, Unaffliliated, Irbid, Irbid, Jordan
[5] Vice Rector for Science and Research, Algebra University, Zagreb, Croatia
[6] Department of Computer Science, College of Computing, Khon Kaen University, Thanon Mittraphap, Thailand
[7] Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon
[8] Department of Applied Data Science, Noroff University College, Kristiansand, Norway
[9] Department of Computer Engineering, Inha University, Incheon, Republic of South Korea

## ABSTRACT

Machine translation has made significant progress in several Indian languages; however, some, known as computationally low-resourced languages, have seen very little work in this field. The Dogri language, which is listed in the 8th Schedule of the Indian Constitution, is one such language. The authors have developed a machine translation system for the Hindi-Dogri language pair in the fixed news domain using three approaches: rule-based machine translation (developed using linguistic rules), statistical machine translation (built using the Moses toolkit), and neural machine translation (developed using neural networks). A comparison of all three approaches is presented in this article. The article also discusses various research challenges identified in each approach used for machine translation. A *corpus* of approximately 0.1 million sentences in the news domain was used to train the *corpus*-based statistical machine translation (SMT) and neural machine translation (NMT) models. The authors also addressed whether NMT produces results equivalent to or better than those of SMT and rule-based machine translation (RBMT). To ensure a comprehensive evaluation, the outputs of all systems were evaluated using two approaches: manual evaluation by language experts and automatic evaluation using standard metrics—Bilingual Evaluation Understudy (BLEU), TER (Translation Edit Rate), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and WER (Word Error Rate). Although RBMT achieved the highest overall scores in both automatic and manual evaluations, expert analysis revealed that translations produced by NMT and SMT exhibited less ambiguity. The study concludes that the

Submitted 21 November 2024
Accepted 26 August 2025
Published 15 October 2025

Corresponding authors
Deepak Prashar,
deepak.prashar@lpu.co.in
Jungeun Kim, jekim@inha.ac.kr

Academic editor
Othman Soufan

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj-cs.3218**



How to cite this article Kumar J, Rakhra M, Dubey P, Prashar D, Mrsic L, Khan AA, Kadry S, Kim J. 2025. Comparative analysis of machine translation for Hindi-Dogri text using rule-based, statistical, and neural approaches. **PeerJ Comput. Sci. 11:e3218 DOI 10.7717/peerj-cs.3218**

performance of SMT and NMT systems are likely to improve further with the availability of larger bilingual parallel corpora.

**Subjects** Algorithms and Analysis of Algorithms, Artificial Intelligence, Natural Language and Speech, Text Mining, Neural Networks

**Keywords** Machine translation, Hindi-Dogri language pair, Low-resourced languages, Neural machine translation (NMT), Statistical machine translation (SMT), Rule-based machine translation (RBMT)

## INTRODUCTION

The technological advancements have enabled digitization in every sphere of life, yet there remains a digital divide due to the language barrier. Every person, regardless of gender, age, or geographical domain, needs access to various kinds of information and applications available for use to make daily tasks easier and time-saving. The Government of India has launched several digital initiatives to provide access to information in regional languages. However, there is still much to be done for low-resourced languages like Dogri. As a result, a large portion of the non-English-speaking population remains dependent on manual sources of information as their primary option. It has been observed that no government website in the state currently offers content in Dogri, despite it being declared an official language of the state in September 2020 (*Government of India, 2020*). One of the hindrances in making content available in local languages is the manual effort required to convert the content into Dogri. This highlights the need for developing state-of-the-art (SOTA) automated machine translation systems. Such systems not only speed up the process but are also cost-effective. It can aid in the translation of various documents such as manuals, newspapers, academic content, literature, and other necessary content in less time and in a cost-effective manner. With intent to develop a state-of-the-art (SOTA) machine translation system (MTS) for the Hindi-Dogri language pair, the authors have worked on the three major approaches of machine translations: a system based on a rule-based approach, statistical MTS and a system based on deep learning models. Machine translation (MT) is a method that uses computer software to translate source language text (such as Hindi) to a target language (such as Dogri) while preserving the original meaning of the source language. Translation poses significant challenges for both human translators and machine translation systems, as it requires proper syntax and semantic knowledge of both languages, but MT has emerged over the past 10 years as a useful tool (*Singh, Kumar & Chana, 2021*) for breaking down barriers to communication in natural language processing. MT methods are generally divided into two categories: rule-based and *corpus*-based approaches. Rule-based methods dominated the field from the inception of MT until the 1990s (*Garje et al., 2016*; *Dubey, 2019*). Rule-based machine translation (RBMT) systems rely on bilingual dictionaries and manually crafted rules to translate source text to target text. In this study, the authors employed the direct approach of rule-based machine translation, which is one of the three main RBMT approaches, alongside the indirect and interlingua methods. The direct approach, also known as the

first generation of machine translation, relies on large dictionaries and word-by-word translation with simple grammatical adjustments. It is designed for specific language pairs, particularly closely related ones, making development easier due to shared grammar and vocabulary. However, this approach is limited to bilingual, unidirectional translation and struggles with ambiguous source texts.

With the emergence of bilingual corpora, *corpus*-based approaches became the dominant approach to convert text from one language to another after the 2000s. Three *corpus*-based MT approaches are commonly used: example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT). EBMT, established in the mid-1980s, operates by retrieving similar sentence pairs from a bilingual *corpus* to translate source texts (*Turcato & Popowich, 2023*). If similar sentence pairs can be retrieved, EBMT algorithms produce high-quality translations. However, EBMT approaches have low translation coverage because bilingual corpora cannot include all the linguistic phenomena of the language pairings.

In 1990, *Brown et al. (1990)* introduced the concept of statistical machine translation, where machines learn translation patterns from the *corpus*, removing the need for human experts to manually define rules. By 1993, this concept was formalized into five progressively complex models now known as the IBM alignment models by *Brown et al. (1993)*. These models established a probabilistic foundation for word alignment and translation, marking a major advancement in the development of machine translation systems.

To conclude, the field of machine translation has undergone considerable development throughout the years, progressing through three major approaches: beginning with rule-based methods (*Garje et al., 2016*; *Dubey, 2019*), moving toward statistical machine translation (SMT) (*Brown et al., 1990*), and ultimately transitioning to neural machine translation (NMT) (*Bahdanau, Cho & Bengio, 2014*; *Mahata et al., 2018*).

In this study, the authors have employed all three approaches for translating Hindi text into Dogri text and analyzed the performance using automatic metrics such as Bilingual Evaluation Understudy (BLEU), Translation Edit Rate (TER), Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Word Error Rate (WER), as well as through manual assessment focusing on adequacy, fluency, and ambiguity. This article is organized into several sections: it begins with an overview of the methodology adopted for developing Hindi-to-Dogri MT systems using RBMTS, SMT, and NMT. This is followed by a description of the datasets used, the experimental setup, and the evaluation criteria. Subsequent sections present the analysis of each MT approach, a comparative analysis of the results, and finally, the conclusions drawn from the study.

## Brief about the languages under study

### *Hindi*

Hindi is one of the two official languages of India. Apart from India, the majority of people in Nepal speak Hindi. It is also a protected language in South Africa and the third official court language in the UAE. It is the fourth most spoken language in the world (*Wikipedia, 2024b*).

**Table 1** Phonetic shifts from Hindi to Dogri: common sound correspondences and examples.

| Hindi phoneme | Dogri phoneme | Use case (Hindi word) | Dogri pronunciation | English meaning |
|---|---|---|---|---|
| घ (gha) | क (ka) | घर (ghara) | कर (kara) | House |
| झ (jha) | च (ca) | झंडा (jhaṃdā) | चंडा (caḍā) | Flag |
| ढ (ḍha) | ट (ṭa) | ढाबा (ṭābā), | टाबा (ṭābā) | Roadside eatery |
| ध (dha) | त (ta) | धन (dhana) | तन (tana) | Wealth |

**Table 2** Semantic shifts in dogri words due to tone changes.

| Dogri word | English meaning | Same Dogri word with tone change | English meaning |
|---|---|---|---|
| कुन (kuna) | Insect | कु'न (ku'na) | Who |
| खल्ल (khalla) | Skin | ख'ल्ल (kha'lla) | Down |
| फड़ (phaḍa) | Catch | फ'ड़ (pha'ḍa) | Boasting |

### *Dogri*

Dogri language is spoken by more than 5 million people in northern India (particularly in Jammu & Kashmir, Himachal Pradesh, and some parts of Punjab) and parts of Pakistan as a Pahari language (*Wikipedia, 2024a*). Dogri got the status of an official language of the Union Territory of Jammu and Kashmir by the Jammu and Kashmir Reorganization Act 2019. Dogri got added to the 8th Schedule of the Indian constitution by 92nd amendment in 2003 and came into effect on Jan 8, 2024. Devanagari script is used for writing both Hindi and Dogri languages (*Gupta, 2004*) from left to right; however, a few characteristics that distinguish Dogri from Hindi are discussed below:

i. **Phonetic differences:**
Some consonants produce different sounds in Dogri compared to Hindi, as illustrated in Table 1.

ii. **Tone and meaning:**
In Dogri, a change in the tone of a word can completely alter its meaning. The apostrophe comma (') is used to represent tone changes, and its placement affects the meaning of words, as shown in Table 2.

iii. **Non-usage of certain Hindi Symbols:**
In Dogri, the Hindi symbols Chandrabindu (ँ) and Visarga (ः) are not used.

iv. **Use of specific letters:**
In Dogri, the letters क्ष, ष, ऋ, and ज्ञ are used exclusively for the transliteration of Sanskrit words.

v. **Indication of extra-long vowels:**
Extra-long vowels are indicated using the sign (ऽ). For example, चनाऽ (canā′)–election, ब्हाऽ (bhā′)–marriage, and ग्रांऽ (grāṃ′) – village.

vi. **Triple consonants:**
In Dogri, some words exhibit the triple use of consonants, such as ननान (nanāna)–

sister-in-law, लग्गग (laggaga)–in use, मन्नना (mannannā)–to agree, सस्स (sassa)–mother-in-law, and बब्ब (babba)–father.

vii. **Nasalization as a phoneme:**
In Dogri, nasalization (ं) functions as a distinct phoneme. The following examples show how nasalization changes the meaning of words: तां (tāṃ)–so, ता (tā)–heat; बांग (bāṃga)–the crowing of a cock, and बाग (bāga)–garden.

## METHODOLOGY

The methodology adopted in this study follows a structured framework comprising four key components to ensure a comprehensive evaluation of Hindi-to-Dogri machine translation systems. First, linguistic resources were collected and prepared, including a rule-based lexicon and a Hindi–Dogri bilingual parallel *corpus* used for training and evaluating SMT and NMT models. Second, three machine translation approaches: RBMTS, SMT, and NMT were developed and implemented to facilitate a comparative study. Third, the outputs of each model were evaluated using both automatic evaluation metrics (BLEU, TER, METEOR, and WER) and expert human linguist evaluation based on standard qualitative measures: adequacy, fluency, and ambiguity. Finally, a comparative analysis was conducted to assess the performance of each model using the combined results from automated and manual evaluations, highlighting their respective strengths and limitations.

## RULE-BASED MACHINE TRANSLATION SYSTEM (RBMTS)

A direct approach of a rule-based machine translation system for Hindi text to Dogri text was developed by creating bilingual dictionaries and a collection of linguistic rules for Hindi and Dogri languages. The grammatical structure of the Hindi language text is transferred into the Dogri language text using these intricate rules. The system's main components include pre-processing (tokenization, normalization, replacing the collocation and proper nouns), lexicon lookup, ambiguity resolution, inflectional analysis followed by the handling of the special cases like Kar, Raha and Laga. Figure 1 depicts the architecture of the system. The system aims to accurately translate Hindi text to Dogri text by upholding the integrity of both languages' grammatical rules. By utilizing these components, the system is able to effectively handle various linguistic complexities and nuances present in the texts.

### Dataset used for RBMTS

A dataset of 22,900 words and phrases was collected under different categories, such as the creation of a Hindi-to-Dogri dictionary, collection of collocation phrases, collection of named entities, and standard words. Dataset details are provided in Table 3 for the development of the rule-based Hindi to Dogri machine translation system.

## DATASET FOR *CORPUS*-BASED MT APPROACHES

The dataset consists of 100,000 Hindi–Dogri parallel sentence pairs, each containing fewer than 80 words. The same *corpus* is used to train both SMT and three different NMT

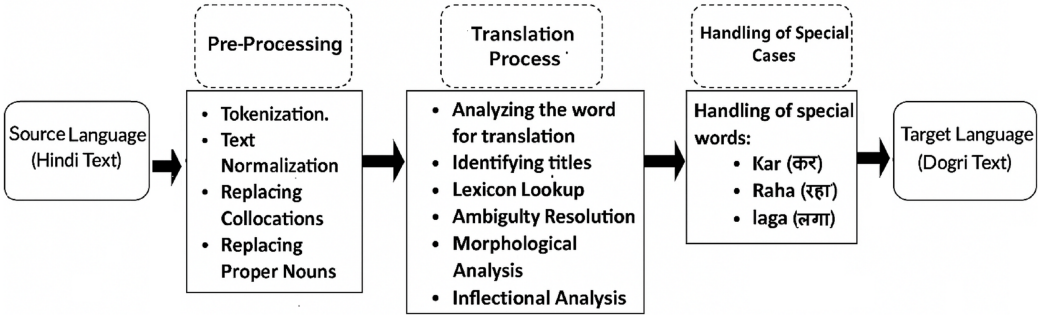


**Figure 1** **Architecture of rule-based machine translation system.**
Full-size DOI: 10.7717/peerj-cs.3218/fig-1

**Table 3** **Hindi-Dogri dataset categorization for rule-based machine translation.**

| S. no | Category | Total word/phrases | Meaning |
|---|---|---|---|
| 1 | Dictionary (Hindi to Dogri) | 18,524 | Each Hindi word is matched with its equivalent Dogri word. |
| 2 | Collocation phrases | 1,834 | Hindi phrases that must be translated as a single unit, not word-by-word. |
| 3 | Named entities | 2,130 | For identification and translation of proper nouns. |
| 4 | Standard words | 412 | Single standard words representing multiple synonyms. |

models, providing a true comparison of the two approaches. The dataset used in the study is collected from a variety of sources, including local and national Hindi newspapers, journals, and news portals, to build the *corpus*. In addition, the authors employed OCR techniques to digitize the Dogri Hindi Conversation Book (डोगरी - हिंदीवार्तालापपुस्तक), a resource published by the Central Hindi Directorate. This book contains common conversational phrases in both Hindi and Dogri (*Central Hindi Directorate of the Government of India, 2018*). Additionally, the authors have gathered various Hindi words and sentences related to collocations, dictionary entries, popular names, and standard vocabulary, and had them translated into Dogri text. The following steps outline the approach adopted by the authors to develop a bilingual Hindi-Dogri parallel *corpus*:

i. Table 4 lists the multiple sources from which the Hindi text was collected.
ii. A rule-based machine translation system (RBMTS) created by *Preeti (2013)* was used to convert the gathered Hindi text into Dogri.
iii. The translated Dogri text was then examined and checked by qualified Dogri linguists to fix any mistakes or inaccuracies brought by the machine translation.
iv. Finally, text alignment was carried out by segmenting the paragraphs into individual sentences and arranging the Hindi and Dogri sentences in a parallel format.

The original work for dataset creation was published in the research article (*Kumar, Rakhra & Dubey, 2022*). To conduct a comparative analysis of all machine translation approaches, 100 randomly selected Hindi sentences from various sources were used to test the RBMTS, SMT, and three NMT models. Table 5 provides statistics on the *corpus* that is

**Table 4** Examples of incorrect Dogri translations produced by RBMTS.

| Source text (Hindi) | English meaning | RBMTS translated Dogri text (Incorrect) | English meaning | Accurate Dogri text |
|---|---|---|---|---|
| उत्तर अमेरिका (uttara amerikā) | North America | **जवाब** अमेरिका (Javāb amerikā) | **Answer** America | उत्तर अमेरिका (uttar amerikā) |
| आम आदमी (āma ādamī) | Common man | **अंब** आदमी (Aanba ādamī) | **Mango** man | आम आदमी (ām ādamī) |
| विजय कुमार (vijaya kumāra) | Vijay Kumar | **जित्त** कुमार (Jitta kumāra) | **Jitta** Kumar | विजय कुमार (vijaya kumāra) |

**Table 5** Examples of incorrect dogri translations produced by RBMTS.

| Hindi word | English meaning | RBMTS translated (Dogri Text) | Possible Dogri translations depending on context | English meaning |
|---|---|---|---|---|
| से (Se) | From | कोला (kolā) | कोला (kolā) | From |
| | | | थमां (thamāan) | By |
| | | | उप्पर (uppara) | Above |
| | | | जेहे (jehe) | Such |
| | | | कन्नै (kannai) | To whom |
| | | | चा (chā) | In |
| | | | दा (dā) | Of |
| | | | शा (shā) | Should |
| दिया (Diyā) | Given | ओड़ेआ (odeā) | ओड़ेआ (odeā) | Given |
| | | | दित्ता (dittā) | Given |
| | | | कीता (kītā) | Done |
| की (Kī) | Of | कीती (kītī) | कीती (kītī) | Done |
| | | | आसेआां (āseāān) | There are |
| | | | दी (dī) | Of |

used for training, validation and testing for the *corpus*-based approaches. The final version of the bilingual *corpus* has been publicly released to the research community and is accessible through the GitHub repository referenced in *Kumar (2024)*.

## Preprocessing of *corpus*

Preprocessing plays a critical role in the development of high-quality MT systems, particularly when dealing with linguistically diverse language pairs such as Hindi and Dogri. Proper preprocessing not only enhances the quality and consistency of the *corpus* but also improves the learning efficiency of both SMT and NMT models. A sample of the final Hindi–Dogri parallel *corpus* is presented in Fig. 2.

In this study, the preprocessing pipeline consisted of several systematic steps designed to clean and standardize the bilingual *corpus* are:

**Sentence length filtering:** Sentences exceeding 80 words in either Hindi or Dogri were removed to eliminate overly long and complex sentence structures that could negatively affect model training.

| Hindi Text | Dogri Text |
|---|---|
| उनकी पहचान अताउल्ला खान पुत्र फ़ज़ा खान, सलीमा बी पत्नी मुश्ताक खान, इरा पठान पुत्री जफर उल्ला खान, अमीरा इश्ताक पुत्री इश्ताक अहमदके रूप में हुई | उंदी पन्छान अताउल्ला खान पुत्र फ़ज़ा खान, सलीमा बी पत्नी मुश्ताक खान, इरा पठान पुत्री जफर उल्ला खान, अमीरा इश्ताक पुत्री इश्ताक अहमद दे रूप च होई |
| डॉक्टरों ने प्राथमिक उपचार अताउल्ला खान व अमीरा इश्ताक को गंभीर हालत के चलते जीएम जम्मू रेफर कर दिया | डाक्टरें प्राथमिक उपचार परेंत अताउल्ला खान ते अमीरा इश्ताक गी गंभीर हालत दे चलदे जीएम जम्मू रेफर करी ओड़ेआ |
| हादसे में मुश्ताक खान पुत्र फजल खान व अताउल्ला खान की मौत हुई | हादसे च मुश्ताक खान जागत फजल खान ते अताउल्ला खान दी मौत होई गेई |
| वह मुश्ताक खान जम्मू कश्मीर पुलिस से सेवानिवृत्त हैं | ओह् मुश्ताक खान जम्मू कश्मीर पुलिस थमां सेवानिवृत्त हा |
| मृतक समेत सभी घायल सलवा मेंढर के निवासी हैं | मृतक समेत सारे जख्मी सलवा मेंढर दे नवासी न |
| जम्मू कश्मीर के ग्रामीण क्षेत्रों में आवासीय भूमि का ड्रोन से सर्वे किया जाएगा | जम्मू कश्मीर दे ग्रांईं क्षेत्रें च आवासीय जमीन दा ड्रोन कन्ने सर्वे कीता जाह्ग |
| इसके बाद लोगों को प्रॉपर्टी कार्ड जारी किए जाएंगे | एह्दे परैंत लोकें गी प्रापर्टी कार्ड जारी कीते जाङन |
| आबादी के सर्वे और प्रॉपर्टी कार्ड की अधिसूचना जारी करने के लिए राजस्व विभाग के आयुक्त सचिव विजय कुमार बिदूरी की तरफ से कस्टोडियन जनरल राजेश शर्मा की अध्यक्षता में छह सदस्यीय कमेटी का गठन किया है | अबादी दे सर्वे ते प्रापर्टी कार्ड दी अधिसूचना जारी करने आस्ते राजस्व मेह्कमें दे आयोग सचिव विजय कुमार बिदूरी दे पास्सेआ कस्टोडियन जनरल राजेश शर्मा दी प्रधानता च छे सदस्यीय कमेटी दा गठन कीता ऐ |
| यह कमेटी 15 दिन में अधिसूचना की पूरी रूपरेखा तैयार कर अपनी रिपोर्ट सरकार को सौंपेगी | एह् कमेटी 15 दिनें अधिसूचना दी पूरी रूपरेखा त्यार करियै अपनी रिपोर्ट सरकार गी सौंपग |
| 31 दिसंबर तक प्रदेश में योजना को लागू किया जा सकता है | 31 दिसंबर तगर प्रदेश च योजना गी लागू कीता जाई सकदा ऐ |
| कमेटी के अध्यक्ष राजेश शर्मा के अनुसार जम्मू कश्मीर के गांवों में आवासीय भूमि पर रह रहे लोगों का सर्वे केंद्रीय सरकार की स्वामित्व योजना के तहत किया जाना है | कमेटी दे प्रधान राजेश शर्मा दे मताबक जम्मू कश्मीर दे गांएं च आवासीय जमीन पर रोहे करदे लोकें दा सर्वे केंद्री सरकार दी स्वामित्व योजना दे तैह्त कीता जाना ऐ |
| योजना के तहत गांवों में आवासीय भूमि के सर्वे ड्रोन के जरिये होगा | योजना दे तैह्त गांएं च आवासीय जमीन दे सर्वे ड्रोन दे जरिए होना ऐ |
| इससे ग्रामीण इलाकों की जमीनों का सीमांकन होगा और ग्रामीण इलाकों में मौजूद घरों के मालिकों के मालिकाना हक का रिकॉर्ड बनेगा | एह्दे कन्ने ग्रांईं लाकैं दियें जमीनें दा सीमांकन होग ते ग्रांईं लाकैं च मजूद घरैं दे मालिकें दे मालकाना हक्क दा रिकार्ड बनग |
| भूमि मालिक संपत्ति कार्ड का इस्तेमाल बैंकों से कर्ज लेने के अलावा अन्य कामों में भी कर सकेंगे | जमीन मालक ज्दाद कार्ड दा इस्तेमाल बैंकें शा दुहार लैने दे अलावा होर केईं कामें च बी करी सकङन |
| स्वामित्व योजना के तहत गांवों की आवासीय भूमि की पैमाइश ड्रोन के जरिए होगी | स्वामित्व योजना दे तैह्त गांवएं दी आवासीय जमीन दी पमेश ड्रोन दे जरिए होग |

**Figure 2** **Sample of bilingual Hindi to Dogri parallel *corpus*.**
Full-size DOI: 10.7717/peerj-cs.3218/fig-2

**Noise removal:** Special characters, duplicated words, incomplete tokens, and extraneous punctuation marks were removed using regular expressions.
**Symbol and HTML tag removal:** Any residual HTML tags and unwanted symbols (like., &, %, @, $) were eliminated to prevent noise during training.
**Normalization of numerals:** Devanagari numerals (०, १, २, …, ९) were replaced with corresponding numerals (0, 1, 2, …, 9) for consistency across the *corpus*.
**Bracketed text elimination:** Any content enclosed within parentheses (), curly braces {}, or square brackets [] was removed to ensure clean sentence structures.
**Manual review and alignment:** After automated cleaning, both the Hindi and Dogri text files were manually reviewed. Sentence pairs were verified and aligned to ensure parallelism and semantic equivalence.

## STATISTICAL MACHINE TRANSLATION SYSTEM

In 1990, *Brown et al. (1990)* introduced the SMT model for machine translation. Building on this foundation, the authors developed an SMT model specifically for Hindi-to-Dogri translation using the Moses toolkit. The Moses toolkit by *Koehn et al. (2007)* trains the translation model using aligned text of both Hindi and Dogri languages. Once the training is complete, the decoder uses beam search to translate the source text to target language text. The beam search algorithm selects the translation with the highest probability. Figure 3 illustrates the architecture of the Hindi-to-Dogri SMT system. SMT analyzes bilingual text corpora to create translation rules, with translation accuracy depending on the quality and size of the bilingual corpora.

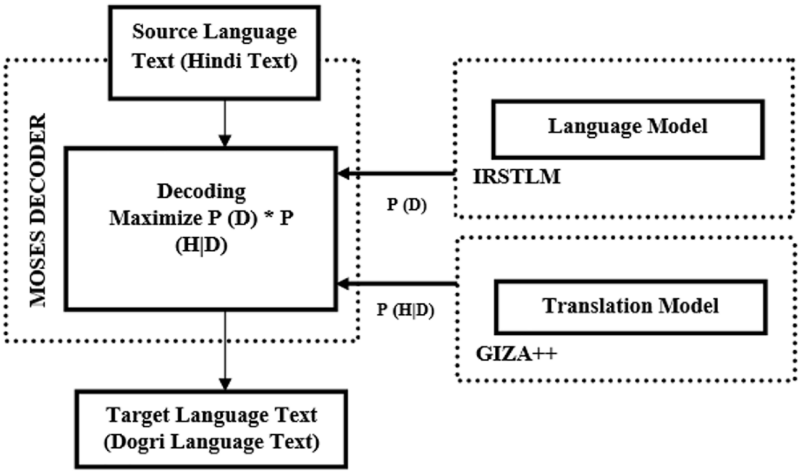


**Figure 3** **Architecture of Hindi to Dogri SMT system.** Full-size DOI: 10.7717/peerj-cs.3218/fig-3

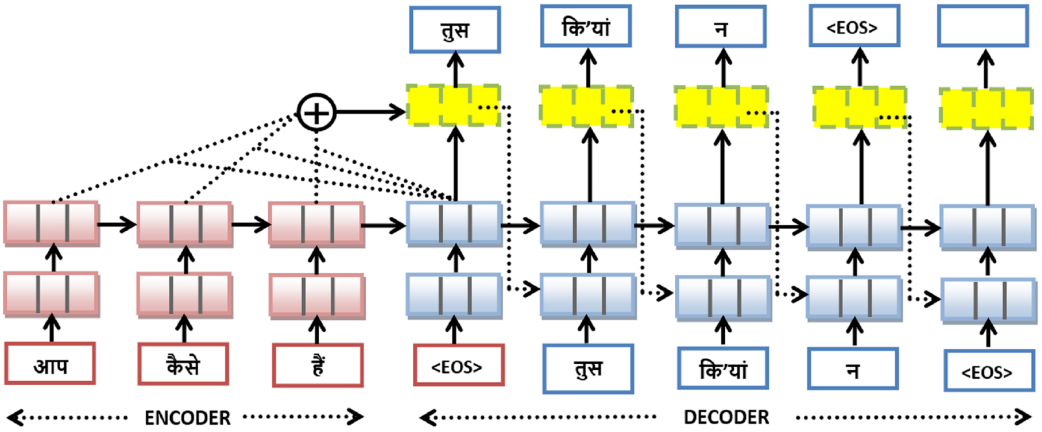


**Figure 4** **Attention-based encoder-decoder architecture for Hindi to Dogri translation.** Full-size DOI: 10.7717/peerj-cs.3218/fig-4

## NEURAL MACHINE TRANSLATION SYSTEM

With the rapid progress of deep learning in domains such as speech recognition and computer vision, researchers began incorporating these techniques into machine translation systems (*Bahdanau, Cho & Bengio, 2014*). Today, the focus has shifted toward more advanced approaches like deep learning-based MTS, which are yielding better results (*Young et al., 2018*). In this study, three deep learning models have been implemented, all based on encoder–decoder architectures with attention mechanisms. Figure 4 serves as a reference to demonstrate the general mechanism of neural machine translation, which the implemented models also follow. On the left side, the encoder (in red) comprises sequential RNN cells—likely LSTM or GRU—that process the input Hindi sentence word-by-word: आप (you), कैसे (how), and हैं (are). On the right, the decoder (in blue) generates the Dogri translation: तुस (you), किंयां (how), न (are), followed by the end-of-sequence token <EOS>. The yellow blocks above the decoder represent attention layers that

dynamically compute context vectors by weighting the encoder outputs. These vectors are aggregated at each decoding step (illustrated by the circle with a "+" symbol) to guide the decoder in producing accurate translations. By training on large parallel corpora and leveraging modern deep-learning frameworks, the model automatically learns which aspects of the input sequence are most informative at each step of output generation, eliminating the need for hand-crafted features and improving both fluency and accuracy.

### Training setup for NMT models

Three different NMT models were developed using the Open-NMT (*Klein et al., 2017*) toolkit, and are referred as recurrent neural network (RNN), bidirectional recurrent neural network with a batch size of 32 (BRNN-32) and bidirectional recurrent neural network with a batch size of 64 (BRNN-64) in this study. The recurrent neural network (RNN) model used a unidirectional LSTM for both encoder and decoder with four layers each, 500 hidden units, an embedding size of 500, a dropout rate of 0.1, learning rate of 1.0, and a batch size of 16. The BRNN-32 model employed a bidirectional LSTM encoder and unidirectional LSTM decoder, with six layers each, 500 hidden units, embedding size of 500, a dropout rate of 0.3, learning rate of 1.0, and a batch size of 32. Similarly, the BRNN-64 model maintained the same architecture as BRNN-32 but increased the batch size to 64. All models were trained for 50 epochs using the SGD optimizer with default gradient clipping of 5 and early stopping set to 4.

All three NMT models were trained on a high-performance workstation running Ubuntu 20.04 LTS. GPU acceleration was provided by an NVIDIA RTX A4000 with 16 GB of VRAM, using NVIDIA driver version 560.35.03 and CUDA version 12.6. The software environment was based on Python and utilized the OpenNMT-py framework, which is built on top of PyTorch and optimized for CUDA-enabled GPU acceleration. The dataset is split into training and validation sets in a ratio of 80:20, with 80% of the dataset used for training, 10% for validation and 10% for testing.

## RESULTS AND FINDINGS

This section analyzes the performance of each MT system and presents key findings, concluding with a comparative analysis of translation results.

### Analysis of the RBMT system

The current rule-based machine translation system relies on a lexicon lookup dictionary containing approximately 22,900 words and phrases. Because of this limited size, many Hindi words and phrases remain untranslated and are directly carried over into the output in Dogri text without any change. This reduces the accuracy of the final output when translating proper nouns, collocations and named entities. Regarding Hindi's polysemous words, such as 'से' (se), 'और' (aur), दिया (diyā), की (kee), *etc*., where the exact translation depends on the context of the discussions, the system generates output with ambiguity. Table 6 shows the output of RBMTS, where the system does not recognize named entities, resulting in incorrect Dogri translations. Table 7 displays a collection of polysemous words

**Table 6** Sources of dataset for the creation of the bilingual Hindi-Dogri parallel *corpus*.

| Hindi text sources | No. of sentences | References | Hindi tokens | Dogri tokens |
|---|---|---|---|---|
| Conversation book on Dogri to Hindi (डोगरी – हिंदीवार्तालापपुस्तक) | 1,802 | http://www.chdpublication.education.gov.in/ebook/b104/html5forpc.html?page=0 | 9,226 | 9,056 |
| Hindi-language online newspapers (such as Amar Ujala, Dainik Jagran, BBC Hindi, Dainik Bhaskar and Hindustan Newspaper) | 20,000 | https://www.amarujala.com/jammu-and-kashmir | 99,905 | 101,139 |
| | 20,000 | https://www.jagran.com/ | 94,914 | 100,232 |
| | 20,000 | https://www.bbc.com/hindi | 92,257 | 93,456 |
| | 20,000 | https://www.bhaskar.com/ | 93,949 | 96,011 |
| | 20,000 | https://www.livehindustan.com/ | 94,836 | 96,114 |
| Dogri name, Hindi collocation, dictionary and standard words | Collected as much as possible Hindi, Dogri words that falls under the category of collocations, dictionary words, popular names and other standard words used both in Hindi as well as Dogri. | | 24,999 | 26,739 |
| **Total** | **101,802** | | **510,086** | **522,747** |

that can take multiple forms depending on the context of the discussion, resulting in ambiguous translation.

The following paragraph presents the translation of Hindi text into Dogri text using RBMT. It contains several incorrect translations of named entities, collocations, and polysemous words. The text marked with strikethrough indicates the incorrect translations produced by the system, while the bold text represents the expected translations. The transliteration of Hindi and Dogri words was carried out using the online transliteration tool available at *Devnagri (2021)*.

### *Hindi text (Input)*

पर्यटन विभाग के निदेशक डॉ. विवेकानंद राय ने सोमवार को बसोहली क्षेत्र के पर्यटन स्थलों का दौरा कर कुछ जरूरी दिशा निर्देश जारी किए हैं। उनहोने साथ में आम लोगों के साथ मुलाक़ात की। इस दौरान उनके साथ डीडीसी अध्यक्ष प्रशांत किशोर, सीओ रोहित सरदाना, सहायक निदेशक विजय शर्मा, बीडीसी अध्यक्ष सुषमा जमवाल और पर्यटन विभाग के अन्य अधिकारी मौजूद रहे। नई योजनाओं पर विचार विमर्श किया। डीडीसी अध्यक्ष ने निदेशक से इलाकों को पर्यटन की दृष्टि से विकसित करने के लिए प्रोजेक्ट बनाने को कहा। इस मंदिर की चारदीवारी करीब एक साल से क्षतिग्रस्त है। इसके अलावा, उन्होंने टूरिज्म रिसेप्शन सेंटर की इमारत का भी जायजा लिया। अंत में पृथ्वी शॉ ने सभी को स्वतंत्रता दिवस की अग्रिम शुभकामनाएं दीं। जहां तीन नदियां गंगा, यमुना और भूमिगत सरस्वती का विलय होता है।

(Paryaḍan vibhāg ke nideshak ḍaॉ. Vivekānanda rāya ne somavār ko basohalī kṣhetra ke paryaḍan sthaloan kā daurā kar kuchh jarūrī dishā nirdesh jārī kie haian।unahone sāth mean ām logoan ke sāth mulāक़ात kī।is daurān unake sāth ḍīḍīsī adhyakṣha prashāanta kishora, sīo rohit saradānā, sahāyak nideshak vijaya sharmā, bīḍīsī adhyakṣha suṣhamā jamavāl aur paryaṭan vibhāg ke anya adhikārī maujūd rahe।naī yojanāoan par vichār vimarsha kiyā।ḍīḍīsī adhyakṣha ne nideshak se ilākoan ko paryaṭan kī dṛuṣhṭi se vikasit karane ke lie projekḍa banāne ko kahā।is mandir kī chāradīvārī karīb ek sāl se kṣhatigrasta hai।isake alāvā, unhoanne ṭūrijma risepshan seanṭar kī imārat kā bhī jāyajā liyā।aanta mean pṛuthvī shaॉ ne sabhī ko svatantratā divas kī agrim shubhakāmanāean dīan।jahāan tīn nadiyāan gangā, yamunā aur bhūmigat sarasvatī kā vilaya hotā hai।).

### *Dogri text output (translated using RBMTS)*

**पर्यटन मैह्कमें** दे निदेशक डा विवेकानंद **राय** नै **सोमवार** गी बसोहली खेतर दे **पर्यटन स्थलें** दा **दौरा करियै** किश **जरूरी दिशा** निर्देश जारी कीते न। **उ'नें** कन्नै **गै आम** लोकें दे कन्नै मुलाकात कीती। इस **दौरान** उं'दे कन्नै डीडीसी प्रधान प्रशांत **किशोर**, सीओ रोहित **सरदाना**, **सहायक** निदेशक **विजय** शर्मा, बीडीसी प्रधान **सुशमा** जमवाल ते **पर्यटन मैह्कमें** दे होर अधिकारी मजूद रेह्। नमीं योजनाओं पर बिचार **विमर्श** कीता। डीडीसी प्रधान नै निदेशक **गी लाके** गी **पर्यटन** दी नज़र **कन्नै** विकसत करने दे आस्तै प्रोजेक्ट बनाने गी आखेआ। इस मंदर **दी** चार-दवारी करीब इक ब'रें शा क्षतिग्रस्त ऐ। एह्दे अलावा, **उ'नें** टूरिज्म रिसेप्शन सेंटर दी अमारत **दी** बी **जांच-पड़ताल कीती**। **खीर** च **पृथ्वी** शा नै सारें गी अजादी दिन **दियां** शुभकामनाओ **दितिया**। जित्थै त्रै नदियां गंगा, जमना ते भूमिगत सरस्वती दा **मेल** होंदा ऐ।

## Analysis of the SMT system

Developing and maintaining rules in a rule-based approach is time-consuming, and transferring them across different domains or languages is a complex task. As a result, scaling rule-based systems for open-domain or multilingual translation is challenging. The SMT model was trained with a parallel *corpus* of approximately 0.1 million sentences, as shown in Table 5. The translation results are generally quite accurate and fluent, barring the translation of rare or unknown words. The system is producing UNK for words which are not part of the training *corpus*. The system managed ambiguity more effectively than RBMTS. The following section, 'Comparative analysis of translation results', supports this observation, showing that the ambiguity score for the SMT model is higher than that of RBMTS.

## Analysis of the NMT system

SMT methods can significantly enhance translation quality; however, they rely on log-linear models that incorporate several manually constructed components, such as the translation model, language model, and reordering model. This often leads to substantial reordering challenges, particularly in distant language pairs. NMT systems are built on neural networks, where each neuron mathematically processes input data. During training, the network is fed bilingual Hindi–Dogri parallel text corpora and adjusts neuron weights based on translation errors. These systems continuously fine-tune themselves, resulting in progressively better performance. Compared to SMT, NMT proves to be more reliable—especially for low-resource languages—due to its superior ability to account for context and generate more natural, human-like translations. This observation is supported by the results summarized in the following section: Comparative Analysis of Translation Results.

## Comparative analysis of translation results

Recent research has documented the differences between various MT systems with respect to the output quality and error types. Some researchers have used automatic evaluation metrics such as TER (*Snover et al., 2006*) and BLEU (*Papineni et al., 2002*) metrics and others have assessed MT systems based on adequacy and fluency through human evaluations of the translation output (*Koehn & Monz, 2006*; *Callison-Burch, Osborne & Koehn, 2006*; *Lavie & Agarwal, 2007*; *White, O'connell & O'mara, 1994*). A few studies have

**Table 7** Sources of dataset for the creation of the bilingual Hindi-Dogri parallel *corpus*.

| Dataset division | Hindi to Dogri text (Sources) | Total no. of Hindi-Dogri parallel sentences | Hindi words | Unique Hindi words | Dogri words | Unique Dogri words |
|---|---|---|---|---|---|---|
| Total *Corpus* used for training, validation and testing of SMT and NMT models | The *corpus* collected from various sources like news papers, books, Standard words, Hindi to Dogri dictionary, Dogri names *etc.* | 100,000 | 771,930 | 67,332 | 777,401 | 66,184 |
| Training *Corpus* | 80% of the total *Corpus* | | | | | |
| Testing *Corpus* | 10% of the total *Corpus* | | | | | |
| Validation *Corpus* | 10% of the total *Corpus* | | | | | |
| Dataset (*Corpus*) for comparative analysis of RBMTS, SMT, and NMT performance | Picked random sentences from News portals | 100 | 1,741 | 156 | 1,742 | 158 |

also combined human evaluation methods with automatic evaluation metrics (AEMs) to provide a more comprehensive analysis (*Jia, Carl & Wang, 2019*). In the present study, the results of all three MT approaches were evaluated using dual framework: expert linguist assessments of adequacy, fluency, and ambiguity, and automated metrics including BLEU, TER, METEOR, and WER. For ambiguity evaluation, the expert linguists employed a custom technique focused specifically on lexical disambiguation. They assessed whether the system correctly interpreted and translated polysemous Hindi words such as: 'से' (sē), which can mean 'by' or 'with' (instrumental); 'और' (aur), meaning either 'and' (conjunctive) or 'more' (comparative); 'दिया' (diyā), which may function as a verb ('gave') or a noun ('light'); and 'की' (kī), which can indicate possession ('of', genitive) or act as an auxiliary/past-tense marker ('did'), among others. As shown in Table 7, these words can have multiple meanings depending on context, making them particularly challenging for machine translation systems. Ambiguity scores were assigned based on the system's ability to accurately translate such context-sensitive polysemous words.

### *Automatic evaluation using standard metrics*

The test dataset was divided into three subsets based on sentence length: SMALL (sentences with fewer than si words), MEDIUM (sentences containing six to 14 words), and LARGE (sentences with 15 or more words). This segmentation aimed to evaluate how sentence length influences translation performance across different models, including rule-based, SMT, and NMT systems. The performance of the Hindi-to-Dogri machine translation systems was quantitatively evaluated using four widely adopted metrics: BLEU, TER, METEOR, and WER. The comparison includes RBMTS, SMT, and various neural machine translation (NMT) models—RNN and bi-directional RNNs (BRNNs) with two batch sizes *i.e.*, 32 and 64. The results revealed significant variations in performance across these subsets as shown in the Table 8. Figures 5–8 present a comparative analysis of the Hindi-to-Dogri translation models based on BLEU, TER, METEOR, and WER scores across small, medium, and large sentence categories.

**Table 8** Evaluation results of Hindi-to-Dogri machine translation using BLEU, TER, METEOR, and WER metrics across different models and sentence lengths.

| Dataset size | Model | BLEU | TER | METEOR | WER |
|---|---|---|---|---|---|
| Small | *RBMTS* | 60.78 | 20.61 | 49.42 | 15.81 |
| | *SMT* | 30.29 | 69.03 | 20.58 | 71.24 |
| | *RNN* | 10.59 | 81.13 | 9.06 | 88.16 |
| | *BRNN (Batch 32)* | 10.27 | 79.72 | 10.02 | 87.56 |
| | *BRNN (Batch 64)* | 8.45 | 84.97 | 7.3 | 92.07 |
| Medium | *RBMTS* | 54.17 | 32.42 | 68.38 | 32.4 |
| | *SMT* | 43.65 | 43.01 | 68.32 | 40.33 |
| | *RNN* | 54.73 | 27.06 | 73.78 | 26.5 |
| | *BRNN (Batch 32)* | 53.99 | 27.45 | 73.75 | 26.99 |
| | *BRNN (Batch 64)* | 52.55 | 28.81 | 71.44 | 28.66 |
| Large | *RBMTS* | 62.08 | 27.32 | 75.14 | 27.59 |
| | *SMT* | 44.73 | 40.17 | 72.56 | 38.83 |
| | *RNN* | 52.72 | 28.35 | 74.68 | 27.69 |
| | *BRNN (Batch 32)* | 54.55 | 27.37 | 75.63 | 26.99 |
| | *BRNN (Batch 64)* | 50.86 | 30.38 | 73.33 | 28.85 |

*BLEU score analysis*

RBMTS attained strong BLEU scores across different sentence lengths (60.78–SMALL, 54.17–MEDIUM, 62.08–LARGE), although the RNN model achieved a slightly higher score (54.73) for MEDIUM-length sentences. Overall, this indicates the strong performance of RBMTS in generating surface-level word matches.

SMT performed reasonably well on MEDIUM and LARGE sentences (43.65 and 44.73, respectively) but struggled with SMALL sentences.

NMT models, especially RNN and BRNNs, showed lower BLEU scores on SMALL sentences but improved substantially on MEDIUM and LARGE datasets. This suggests neural models are more effective in handling complex, longer sentences where context plays a greater role.

*TER score analysis*

RBMTS showed the lowest TER (*i.e.*, fewer edits needed) for SMALL and LARGE sentences, demonstrating fluency and accuracy in simpler structures.

NMT models had higher TER for SMALL sentences, likely due to overfitting or difficulty handling short context.

For MEDIUM and LARGE sentences, RNN and BRNN-32 achieved better TER than SMT, with BRNN models showing slightly better edit efficiency, especially on LARGE sentences.

*METEOR score analysis*

RBMTS again led in SMALL and LARGE sentences, but NMT models (RNN and BRNNs) outperformed both RBMTS and SMT in the MEDIUM sentences, where semantic understanding is more important.

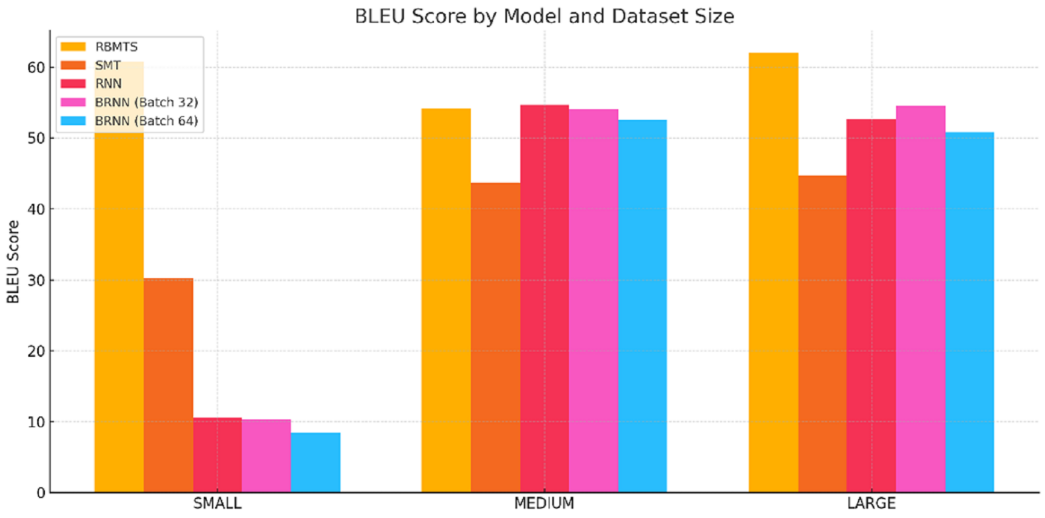


**Figure 5** **BLEU score comparsion of Hindi-to-Dogri translation models across small, medium, and large sentences categories.** Full-size DOI: 10.7717/peerj-cs.3218/fig-5

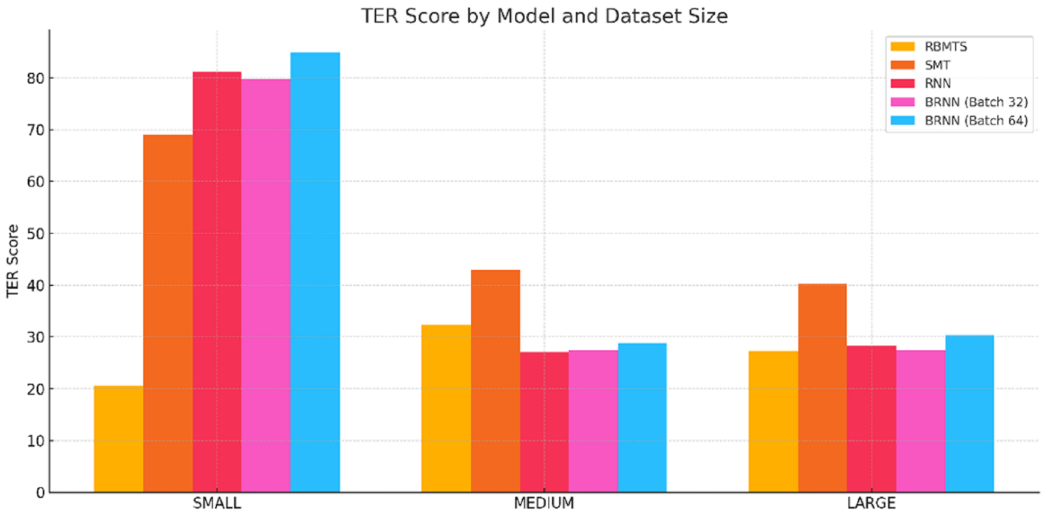


**Figure 6** **TER score comparsion of Hindi-to-Dogri translation models across small, medium, and large sentences categories.** Full-size DOI: 10.7717/peerj-cs.3218/fig-6

Interestingly, BRNN-32 achieved the highest METEOR score for LARGE sentences (75.63), indicating its superior ability to maintain semantic equivalence and word alignment in longer contexts.

*WER score analysis*

RBMTS had the lowest WER across all sentence's sizes, reflecting fewer word-level errors and better literal translation.

SMT and NMT models, particularly RNN and BRNNs, had significantly higher WER in SMALL sentences, reinforcing the observation that short sentences pose challenges for context-dependent models.

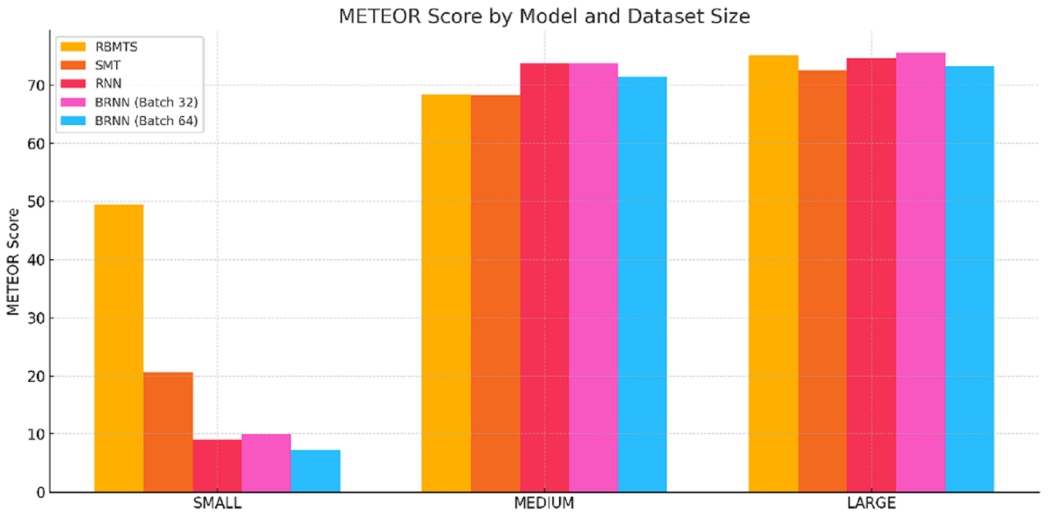


**Figure 7** **METEOR score comparsion of Hindi-to-Dogri translation models across small, medium, and large sentences categories.** Full-size DOI: 10.7717/peerj-cs.3218/fig-7

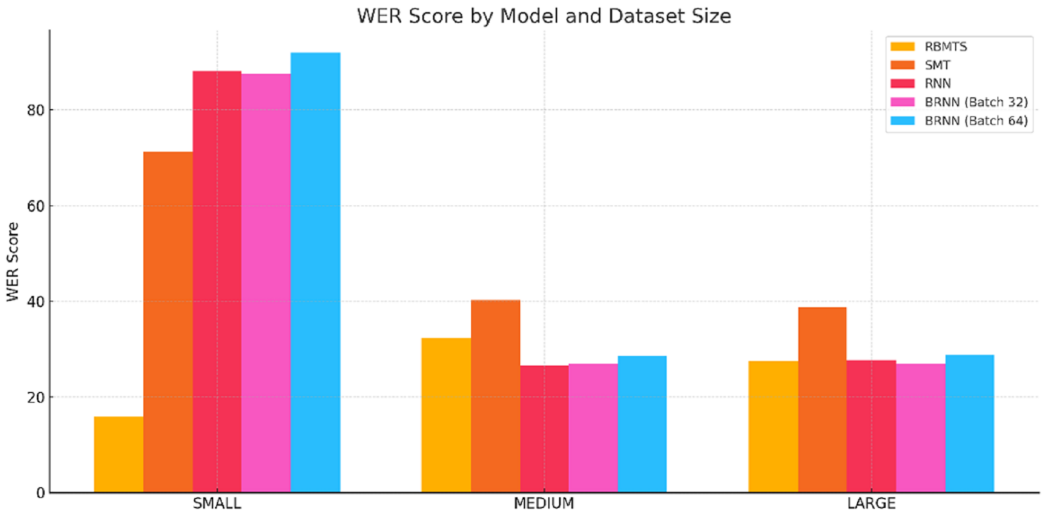


**Figure 8** **WER score comparsion of Hindi-to-Dogri translation models across small, medium, and large sentences categories.** Full-size DOI: 10.7717/peerj-cs.3218/fig-8

In MEDIUM and LARGE sentences, BRNN-32 exhibited the best WER among neural models, indicating its improved reliability in realistic sentence lengths.

### *Human evaluation of results*

For expert manual evaluation, an Excel spreadsheet was compiled containing Hindi source sentences along with their corresponding Dogri translations generated by various models. Three native Dogri-speaking professional linguists evaluated the translations based on three metrics: Adequacy, which measures how well the translated text preserves the meaning of the original Hindi input; fluency, which assesses the grammatical correctness and naturalness of the Dogri output; and ambiguity, which evaluates the clarity of the

**Table 9 Evaluation results of Hindi-to-Dogri machine translation using BLEU, TER, METEOR, and WER metrics across different models and sentence lengths.**

| Score | Adequacy | Fluency | Ambiguity |
|---|---|---|---|
| 5 | Dogri output fully preserves meaning of Hindi sentence. | Native-level Dogri; no grammatical or stylistic issues. | Completely clear; no ambiguity or confusion |
| 4 | Minor details are missing but meaning is mostly intact. | Small errors but very natural overall. | Slightly unclear in rare cases, but meaning is generally obvious. |
| 3 | Some parts are missing or altered; overall idea is understandable. | Understandable but contains awkward or incorrect phrases. | Some ambiguity present; meaning requires interpretation. |
| 2 | Only a small part of the meaning is conveyed. | Hard to read; poor grammar. | Significant ambiguity; multiple possible interpretations. |
| 1 | Translation is incorrect or meaningless. | Broken Dogri; hard or impossible to understand. | Highly ambiguous or completely unclear meaning. |

translation and whether it is prone to multiple interpretations or confusion. Each of these metrics adequacy, fluency, and ambiguity was rated using a five-point Likert scale, where a score of 5 indicates excellent quality (*i.e.*, complete meaning preservation, flawless grammar, and clear interpretation), while a score of 1 indicates poor quality (*i.e.*, significant meaning loss, serious grammatical errors, or high ambiguity).

Table 9 outlines the rating scale used for adequacy, fluency, and ambiguity. Higher the value means better results. The final scores were calculated by averaging the ratings across all evaluators and sentences, as shown in the equation:

- N be the total number of evaluated sentences
- M be the total number of human evaluators
- s_{i,j} denote the adequacy score assigned by evaluator j to sentence i
- f_{i,j} denote the fluency score assigned by evaluator j to sentence i
- a_{i,j} denote the ambiguity score assigned by evaluator j to sentence i

*Metric score (M_X)*

For any evaluation metric X (where X ∈ {A, F, A_m}, corresponding to Adequacy, Fluency, and Ambiguity, respectively), let:

- x_{i,j} denote the score given by evaluator j to sentence i for metric X.

The score for metric X is then calculated as: $M_x = (1/N)\Sigma_{i=1}^{n}[(1/M)\Sigma_{j=1}^{m} x_{i,j}]$

A set of 100 Hindi sentences of varying lengths was collected for manual evaluation. Using the equation $M_X$, the evaluation scores were calculated and are presented in Table 10. The comparative manual evaluation scores, analyzed in the following section, are presented in Fig. 9.

**RBMTS:** outperforms other systems in adequacy (2.90) and fluency (2.80), suggesting it can better preserve meaning and grammatical structure. However, its ambiguity score (2.20) is the lowest, indicating that while accurate, its outputs may lack clarity or sound overly rigid due to rule limitations.

**Table 10** Evaluation results of Hindi-to-Dogri machine translation using BLEU, TER, METEOR, and WER metrics across different models and sentence lengths.

| MT system | Adequacy score | Fluency score | Ambiguity score |
|---|---|---|---|
| RBMTS | 2.9 | 2.8 | 2.2 |
| SMT | 2.2 | 2.2 | 2.6 |
| RNN | 2.68 | 2.6 | 2.8 |
| BRNN (Batch 32) | 2.71 | 2.65 | 2.9 |
| BRNN (Batch 64) | 2.58 | 2.58 | 2.7 |

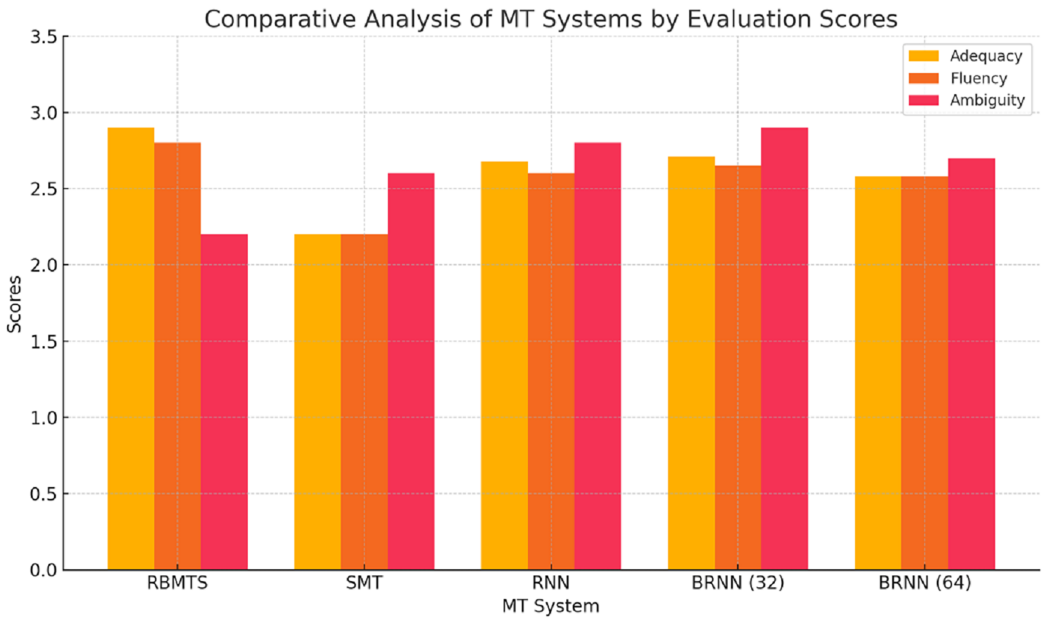


**Figure 9** Comparative analysis of MT systems by manual evaluation scores.
Full-size DOI: 10.7717/peerj-cs.3218/fig-9

**SMT:** receives the lowest scores in adequacy and fluency (2.20 each), showing it often fails to retain complete meaning and produce natural sentences. Its relatively higher ambiguity score (2.60) indicates that its output is somewhat clearer, possibly due to over-simplified phrasing.

**RNN:** improves upon SMT in all areas, particularly in ambiguity (2.80), highlighting that it produces more contextually coherent and less confusing translations. However, it still falls slightly behind RBMTS in adequacy and fluency.

**BRNN-32:** achieves the highest ambiguity score (2.90) and slightly edges out RNN in adequacy and fluency, indicating that it provides a better balance of meaning retention and sentence clarity. The bidirectional context handling improves its translation quality noticeably.

**BRNN-64**: shows comparable results but with a slight dip in all scores compared to Batch 32. This may reflect that larger batch sizes don't always translate to better sentence-level precision, possibly due to convergence smoothing in training.

### *Sample output*

Among the 100 sample sentences chosen for manual expert evaluation, three examples are provided to showcase how each model translates the given Hindi text. The accuracy of each output is evaluated by comparing it with a reference translation in Dogri:

**Hindi Text 1**: ताकि लोगों को सुविधा मिल सके (tāki logoan ko suvidhā mil sake)
***Dogri Reference Text 1:*** *तांजे लोकें गी सुविधा थ्होई सकै (tāanje lokean gī suvidhā thhoī sakai)*
**Dogri Text 1 (RBMTS):** तांजे लोकें गी सुबधा मिल्ल सकै (tāanje lokean gī subadhā milla sakai)
**Dogri Text 1 (SMT):** तांजे लोकां गी सुबधा थ्होई सकै (tāanje lokāan gī subadhā thhoī sakai)
**Dogri Text 1 (RNN):** तांजे लोकें गी सुबधा मिली सकै (tāanje lokean gī subadhā milī sakai)
**Dogri Text 1 (BRNN-32):** तांजे लोकें गी सुबधा थ्होई सकै (tāanje lokean gī subadhā thhoī sakai)
**Dogri Text 1 (BRNN-64):** तांजे लोकें गी सुबधा मिली सकै (tāanje lokean gī subadhā milī sakai)
**Hindi Text 2:** मैं बस से यात्रा कर रहा हूँ (maiṅ bas sē yātrā kar rahā hūṅ)
***Dogri Reference Text 2:*** *में बस्स थमां जात्तरा करा दा आं (mean bassa thamāan jāttarā karā dā āan)*
**Dogri Text 2 (RBMTS):** में बस्स शा जातरा करा दा आं (mean bassa shā jātarā karā dā āan)
**Dogri Text 2 (SMT):** में बस्स पर यात्रा करा दा आं (mean bassa par yātrā karā dā āan)
**Dogri Text 2 (RNN):** में बस्स थमां जात्तरा करा दा आं (mean bassa thamāan jāttarā karā dā āan)
**Dogri Text 2 (BRNN-32):** मैं बस थमां जात्तरा करै करदे आं (maian bas thamāan jāttarā karai karade āan)
**Dogri Text 2 (BRNN-64):** मैं बस थमां जात्तरा करा करनां करनी आं (maian bas thamāan jāttarā karā karanāan karanī āan)
**Hindi Text 3:** रक्षाबंधन का पर्व भाई और बहन के बीच अट्टू प्रेम का प्रतीक है (rakṣhābandhan kā parva bhāī aur bahan ke bīch aṭṭū prem kā pratīk hai)
***Dogri Reference Text 3:*** *रक्खड़ी दा पर्व भ्राऽ ते भैन च अट्टू प्यार दा प्रतीक (rakkhaḍī dā parva bhrā' te bhain ch aṭṭū pyār dā pratīk)*
**Dogri Text 3 (RBMTS):** रक्खड़ी दा पर्व भ्राऽ ते भैन दे बिच्च अट्टू प्यार दा नशानी ऐ (rakkhaḍī dā parva bhrā' te bhain de bichcha aṭṭū pyār dā nashānī ai)
**Dogri Text 3 (SMT):** रक्खड़ी दा पर्व भ्राऽ ते भैन च अट्टू प्यार दा प्रतीक ऐ (rakkhaḍī dā parva bhrā' te bhain ch aṭṭū pyār dā pratīk ai)
**Dogri Text 3 (RNN):** रक्खड़ी दा पर्व भ्राऽ ते भैन च अट्टू प्यार दा प्रतीक ऐ (rakkhaḍī dā parva bhrā' te bhain ch aṭṭū pyār dā pratīk ai)
**Dogri Text 3 (BRNN-32):** रक्खड़ी दा पर्व भ्राऽ ते भैन च अट्टू प्यार दा प्रतीक ऐ (rakkhaḍī dā parva bhrā' te bhain ch aṭṭū pyār dā pratīk ai)
**Dogri Text 3 (BRNN-64):** रक्खड़ी दा पर्व भ्राऽ ते भैन च अट्टू प्यार दा नशानी ऐ (rakkhaड़ī dā parva bhrā' te bhain ch aṭṭū pyār dā nashānī ai)

## CONCLUSION

This study presents a comparative evaluation of three machine translation (MT) approaches—RBMTS, SMT, neural (RNN and BRNN with batch sizes of 32 and 64)—for

Hindi-to-Dogri translation. Each method exhibits distinct strengths and limitations, making the choice of MT system highly dependent on language pair characteristics, domain specificity, data availability, and intended application. From the automatic evaluation, it was observed that RBMTS performs optimally for short and structurally simple sentences due to its reliance on handcrafted linguistic rules. However, as sentence complexity and length increase, neural models—particularly the BRNN with batch size 32—demonstrate superior performance across BLEU, METEOR, TER, and WER metrics. This model captures semantic and contextual nuances more effectively, offering a balanced solution for varying sentence lengths. SMT, while more robust than basic NMT models, consistently lags behind both RBMTS and BRNNs models. The human evaluation further supports these findings. RBMTS retains structural and grammatical fidelity and demonstrates high adequacy in meaning preservation. However, it introduces semantic rigidity, occasionally leading to ambiguity. In contrast, data-driven neural models—especially BRNN-32—yield more fluent and natural Dogri translations, showing better clarity and improved handling of ambiguous phrases. Among all evaluated systems, BRNN-32 achieves the best overall human ratings for fluency and ambiguity while maintaining competitive adequacy scores. A common limitation across all systems was their inability to handle out-of-vocabulary (OOV) or unknown words. In the case of RBMTS, such words were simply retained in the translation without being translated. Addressing this issue through advanced techniques like subword units, back-translation, or named entity recognition could substantially improve translation. Overall, while RBMTS remains a strong contender for low-resource, syntax-aligned language pairs like Hindi-Dogri, the results indicate that neural models—especially BRNNs—are better suited for scalable, flexible translation systems. With access to larger and more diverse parallel corpora, SMT and NMT models are expected to significantly improve in both adequacy and fluency. Future work will explore hybrid MT architectures that leverage the linguistic precision of RBMTS with the contextual depth of NMT, aiming to develop robust systems capable of delivering high-quality translations across diverse linguistic contexts.

## ADDITIONAL INFORMATION AND DECLARATIONS


### Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00277907) and by the Technology Innovation Program (No. 20022899) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:
National Research Foundation of Korea (NRF).
Korea government (MSIT): RS-2023-00277907.
Ministry of Trade, Industry & Energy (MOTIE, Korea): 20022899.

## Competing Interests

The authors declare that they have no competing interests.

## Author Contributions

- Joginder Kumar conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Manik Rakhra conceived and designed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Preeti Dubey performed the experiments, prepared figures and/or tables, and approved the final draft.
- Deepak Prashar analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Leo Mrsic analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Arfat Ahmad Khan analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Seifedine Kadry analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Jungeun Kim analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

## Data Availability

The following information was supplied regarding data availability:

The data is available at GitHub and Zenodo:

- https://github.com/jpbadgal/Dataset.

- jpbadgal. (2025). jpbadgal/Dataset: Hindi–Dogri Bilingual Parallel *Corpus* (Dataset) (v1.0). Zenodo. https://doi.org/10.5281/zenodo.17038614.

The code is available in the Supplemental File.

## Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3218#supplemental-information.

# REFERENCES

**Bahdanau D, Cho K, Bengio Y. 2014.** Neural machine translation by jointly learning to align and translate. ArXiv DOI 10.48550/arXiv.1409.0473.

**Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty JD, Mercer RL, Roossin PS. 1990.** A statistical approach to machine translation. *Computational Linguistics* **16(2)**:79.

**Brown PE, Della Pietra VJ, Della Pietra SA, Mercer RL. 1993.** The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* **19(2)**:263.

**Callison-Burch C, Osborne M, Koehn P. 2006.** Re-evaluating the role of BLEU in machine translation research. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg: ACL, 249–256.

**Central Hindi Directorate of the Government of India. 2018.** Dogri-Hindi conversation book. *Available at http://www.chdpublication.education.gov.in/ebook/b104/html5forpc.html?page=0* (accessed 26 April 2024).

**Devnagri AI. 2021.** Devnagri transliteration demo. *Available at https://transliteration.devnagri.com/* (accessed 12 March 2023).

**Dubey P. 2019.** The Hindi to Dogri machine translation system: grammatical perspective. *International Journal of Information Technology* **11(1)**:171–182 DOI 10.1007/s41870-018-0085-4.

**Garje GV, Bansode A, Gandhi S, Kulkarni A. 2016.** Marathi to English sentence translator for simple assertive and interrogative sentences. *International Journal of Computer Applications* **138**:42 DOI 10.5120/ijca2016908837.

**Government of India. 2020.** The Jammu and Kashmir Official languages act, 2020. *Available at https://www.indiacode.nic.in/bitstream/123456789/15512/1/A2020_23.pdf*.

**Gupta V. 2004.** *Dogri Vyakaran*. Jammu: Jammu and Kashmir Academy of Art, Culture and Language.

**Jia Y, Carl M, Wang X. 2019.** Post-editing neural machine translation versus phrase-based machine translation for English-Chinese. *Machine Translation* **33(1–2)**:9–29 DOI 10.1007/s10590-019-09229-6.

**Klein G, Kim Y, Deng Y, Senellart J, Rush A. 2017.** OPENMT: open-source toolkit for neural machine translation. In: Bansal MJH, ed. *Proceedings of ACL 2017, System Demonstrations*. Stroudsburg: ACL, 67–72. *Available at https://aclanthology.org/P17-4012/*.

**Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Mit CM, Zens R, Aachen R, Dyer C, Bojar O, Cornell EH. 2007.** Moses: open source Toolkit for statistical machine translation ITC-irst 2. In: *ACL'07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Stroudsburg: ACL.

**Koehn P, Monz C. 2006.** Manual and automatic evaluation of machine translation between European languages 1 evaluation framework. *In: Proceedings on the Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics, 102–121. *Available at https://aclanthology.org/W06-3114/*.

**Kumar J. 2024.** Bilingual Hindi-to-Dogri Parallel corpus (Dataset) (School of Computer Applications, Lovely Professional University, Punjab, India). GitHub. *Available at https://github.com/jpbadgal/Dataset* (accessed 26 April 2024).

**Kumar J, Rakhra M, Dubey P. 2022.** Bilingual parallel corpora: a major resource for developing computational tools for automatic processing of Hindi-Dogri language pair. In: *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. Piscataway: IEEE, 1–6 DOI 10.1109/ICRITO56286.2022.9964875.

**Lavie A, Agarwal A. 2007.** Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: *StatMT'07: Proceedings of the Second Workshop on Statistical Machine Translation*.

**Mahata SK, Mandal S, Das D, Bandyopadhyay S. 2018.** SMT vs NMT: a comparison over Hindi & Bengali simple sentences. ArXiv DOI 10.48550/arXiv.1812.04898.

**Papineni K, Roukos S, Ward T, Zhu W-J. 2002.** BLEU: a method for automatic evaluation of machine translation. In: *ACL'02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg: ACL.

**Preeti D. 2013.** *Study and development of machine translation system from Hindi language to Dogri language an important tool to bridge the digital divide*. Thesis. University of Jammu, Jammu, India. *Available at http://hdl.handle.net/10603/78191*.

**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 DOI 10.1007/s11831-020-09449-7.

**Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J. 2006.** A study of translation edit rate with targeted human annotation. In: *AMTA, 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation*, 223–231.

**Turcato D, Popowich F. 2023.** What is example-based machine translation? In: Carl M, Way A, eds. *Recent Advances in Example-Based Machine Translation. Text, Speech and Language Technology*. Vol. 21. Dordrecht: Springer DOI 10.1007/978-94-010-0181-6_2.

**White JS, O'connell T, O'mara F. 1994.** The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In: *Proceedings of the First Conference of the Association for Machine Translation in the Americas. Available at https://aclanthology.org/1994.amta-1.25/*.

**Wikipedia. 2024a.** Dogri language. *Available at https://en.wikipedia.org/wiki/Dogri_language* (accessed 10 July 2024).

**Wikipedia. 2024b.** Hindi. *Available at https://en.wikipedia.org/wiki/Hindi* (accessed 10 June 2024).

**Young T, Hazarika D, Poria S, Cambria E. 2018.** Recent trends in deep learning based natural language processing [Review Article]. *IEEE Computational Intelligence Magazine* **13(3)**:55–75 DOI 10.1109/MCI.2018.2840738.