**International Journal of Advances in Data and Information Systems**
Vol. 5, No. 2, October 2024, pp. 101~112
ISSN: 2721-3056, DOI: 10.59395/ijadis.v5i2.1323


# Indonesian to Bengkulu Malay Statistical Machine Translation System

**Bella Okta Sari Miranda*[1], Herman Yuliansyah[2], Muhammad Kunta Biddinika[3]**
[1,3]Master Program of Informatics, Universitas Ahmad Dahlan, Indonesia
[2]Department of Informatics, Universitas Ahmad Dahlan, Indonesia

**Article Info**

***Article History:***

Received May 09, 2024
Revised Jul 21, 2024
Accepted Oct 30, 2024

***Keywords:***

Bengkulu Malay Language
BLEU
Indonesian Language
Parallel Corpus
Statistical Machine Translation

**ABSTRACT**

Machine translation is an automatic tool that can process language translation from one language to another. This research focuses on developing Statistical Machine Translation (SMT) from Indonesian to Bengkulu Malay and evaluating the quality of the machine translation output. The training and testing data consist of parallel corpora taken from Bengkulu Malay dictionaries and online resources for Indonesian corpora, with a total of 5261 parallel sentence pairs. Several steps are performed in SMT. The initial step is preprocessing, aimed at preparing the parallel corpus. After that, a training phase is conducted, where the parallel corpus is processed to build language and translation models. Subsequently, a testing phase is carried out, followed by an evaluation phase. Based on the research results, various factors influence the quality of SMT translation output. The most important factor is the quantity and quality of the parallel corpus used as the foundation for developing translation and language models. The machine translation output is automatically evaluated using the Bilingual Evaluation Understudy (BLEU), indicating accuracy values observed when using 500 sentences, 1500 sentences, 2500 sentences, 4000 sentences, and 5261 sentences are 80.56%, 90.86%, 92.50%, 92.91%, and 94.48% respectively.



***Corresponding Author:***

Bella Okta Sari Miranda
Faculty of Industrial Technology,
Universitas Ahmad Dahlan,
Jl. Ringroad Selatan, Kragilan, Tamanan, Banguntapan, Bantul, Special Region of Yogyakarta
Email: mirandabella1110@gmail.com

## 1. INTRODUCTION

Language plays a very significant role in human life. Through language, humans can convey their feelings and thoughts to others. Communication and interaction among individuals become possible with the existence of language. Humans, by their nature, are social beings who cannot exist without the involvement of other humans in their lives. When in the process of socializing, humans require a means of communication, which is manifested through language [1][2]. The Malay Bengkulu language is commonly used by the inhabitants of the Bengkulu province. This language is included in the West Austronesian language group, like other languages of the Malay Archipelago. However, the Malay Bengkulu language has distinctive differences from other West Austronesian languages [3][4]. The Bengkulu province has nine local languages that are commonly used in the conversations of its community. These languages include Lembak, Pekal, Serawai, Rejang, Bintuhan, Malay Bengkulu, Muko-Muko, Enggano, and Pasemah. Until now, all of these local languages are still preserved and used by the local community as a means of communication in their daily lives [5].

The city of Bengkulu has experienced rapid growth and has become one of the most densely populated areas in the Bengkulu province. Many residents come to the city of Bengkulu from various regions including Java, South Sumatra, West Sumatra, Lampung, and other areas. They bring with them cultural wealth and customs from their respective regions, which then create a multicultural and diverse environment in the city of Bengkulu. The presence of various ethnic and cultural groups also contributes to the diversity of languages in the city, with residents speaking various local languages as well as Indonesian as the official language [6][7].

The shift in language usage in Bengkulu society, especially the increasing use of Indonesian in daily conversations, can result in a decline in the use of local languages, particularly Malay Bengkulu. This can impact the preservation of the cultural and linguistic identity of the Bengkulu community. Language and cultural diversity are valuable assets to a society. Therefore, it is important to take proactive steps to preserve local languages, especially Malay Bengkulu. To avoid the threat of extinction of local languages, the preservation of these languages becomes crucial.

Technology can be an effective tool in introducing, preserving, and strengthening the use of local languages [8]. One promising potential in technology development is machine translation. This machine serves as an effective solution to the challenges that arise in the language translation process. With advancements in artificial intelligence and natural language processing technology, machine translation is increasingly capable of producing accurate and high-quality translations, enabling smoother and more efficient cross-language communication. With the existence of machine translation, limitations that may have previously hindered communication between languages can now be overcome more easily. In processing language from humans and computers, a Natural Language Processing (NLP) system is required. NLP is one of the branches of computer science, artificial intelligence, and linguistics that focuses on the relationship between computers and natural human language, such as Indonesian or other languages [9].

Machine translation is an automatic tool that can process language translation from one language to another. The purpose of creating machine translation is to facilitate communication interaction between people speaking different languages. Some researchers have previously conducted research on statistical machine translation for local languages. Some of these studies include Phrase-Based Statistical Machine Translation for Javanese-Indonesian [10], Statistical Machine Translation for Lampung Api Dialect to Indonesian Language [11], Statistical Machine Translation Muna To Indonesia Language [12], Minang and Indonesian Phrase-Based Statistical Machine Translation [13], Statistical Machine Translation Dayak Language – Indonesia Language [14].

This research utilizes a machine translator that employs a statistical approach or SMT to translate text from Indonesian to Bengkulu Malay. SMT is an approach to Machine Translation characterized by the use of machine learning methods. The statistical approach used is based on the concept of probability. The higher the probability value, the better-formed the translated sentences are. One advantage of using statistical machine translation is that with a larger corpus, the machine can understand the "context" of frequently occurring phrases, thereby producing more accurate translations [15][16].

The use of SMT marks an important initial step in the development of machine translators, especially for Malay Bengkulu Language. This step demonstrates concrete efforts in preserving local languages and preventing language extinction by leveraging technology. By implementing SMT for Malay Bengkulu Language, it is hoped that accessibility to this language can be expanded, facilitating interaction and information exchange among communities that use it. This not only helps to preserve the continuity and cultural richness of Malay Bengkulu Language but also enriches the overall linguistic ecosystem. Therefore, the development of machine translators focusing on local languages like Malay Bengkulu Language has broad and significant impacts in the context of preserving the heritage of local language and culture.

The aim of this research is to develop an effective and reliable statistical machine translator for translating from Indonesian to Malay Bengkulu Language. Additionally, this research also aims to develop the parallel corpus, with a target of reaching 5261 data points. The development of this parallel corpus is expected to significantly contribute to improving the quality of the machine

translator. By having a larger and more representative parallel corpus, the machine translator will have access to more data and contexts, which will enhance its ability to produce accurate and precise translations in Malay Bengkulu Language.

## 2. RESEARCH METHOD

In general, the development of statistical machine translation involves several sequential steps. The first stage involves preparing the main data, primarily by compiling parallel corpora. Next, these parallel corpora undergo a series of processes, starting from the preprocessing stage, which involves cleaning and preparing the data for training. Following that, it proceeds to the training stage, where the machine translation model is provided with data and learns the language. After the model is trained, the decoding process is carried out to translate new texts, and an evaluation stage is used to assess the system's performance. Furthermore, the methodology for statistical machine translation research from Indonesian to Bengkulu Malay is detailed in figure 1 attached below.

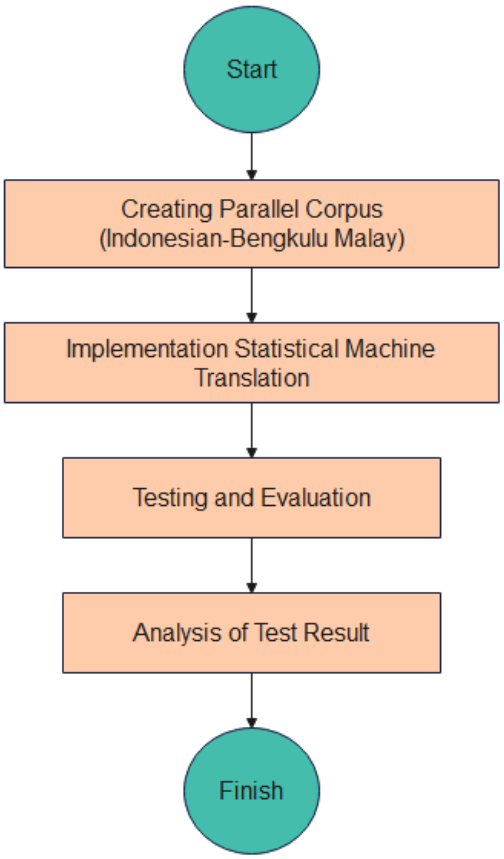


Figure 1. Research Flow

### 2.1 Creating Parallel Corpus

In this reseach, the data used originates from text documents in Bengkulu Malay taken from the Bengkulu Malay dictionary. The data collection process involved efforts to obtain Bengkulu Malay sentences that already have translations in Indonesian. To facilitate computational research, a parallel corpus covering both languages is required. Parallel corpus is a collection of two corpora containing sentences in one language and their translations into another language [17]. Manual typing processes were undertaken to create this parallel corpus, resulting in 5261 pairs of sentences successfully collected in both Indonesian and Bengkulu Malay. By utilizing this extensive parallel corpus, the research aims to achieve a deeper understanding of the translation process between these two languages in a computational context.

Table 1. Example Dataset

| Indonesian | Bengkulu Malay |
|---|---|
| Anak saya kemarin perempuan | Anak ambo kemaren betino |
| Sepertinya orang-orang ini mengambil tanah | Caknyo orang-orang iko ngambik tanah caknyo |
| Hari senin nanti saya mulai sekolah lagi | Ari senayan kelak ambo mulai sekolah lagi |
| Kaki saya terbentur dengan batu besar | Kaki ambo terantuk dekek batu besak |
| Kita bertemu dipasar minggu aja | Kito bersuo dipasar minggu ajo |

## 2.2 Implementation Statistical Machine Translation

The first step in developing a Statistical Machine Translation system from Indonesian to Bengkulu Malay involves designing the system architecture. This design phase entails determining the structure and main components of the statistical translation engine. In detail, a description of the system architecture can be found in the attached image below. This process serves as a crucial foundation for further implementing and optimizing the performance of the statistical machine translation system. Can be seen in Figure 2 below.

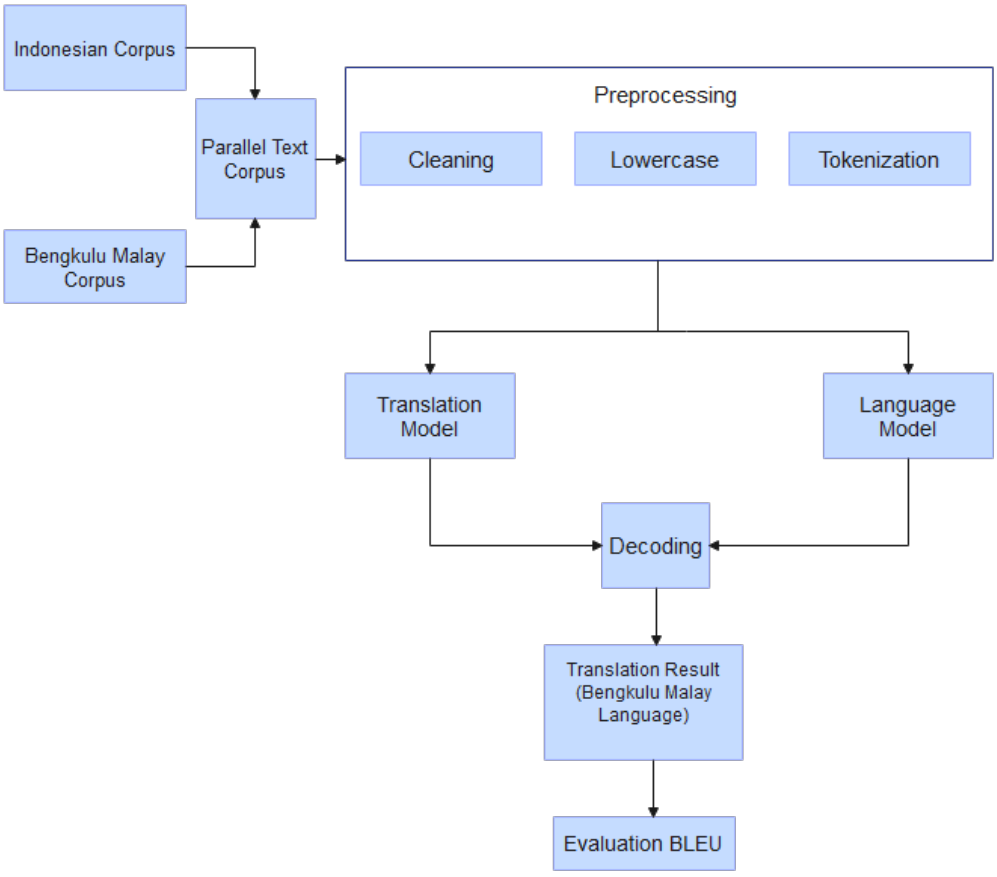


Figure 2. Architecture of the Statistical Machine Translation Indonesian-Bengkulu Malay

Figure 2 depicts the architecture of a statistical machine translation system involving several steps, such as preprocessing, language model formation, translation model training, decoding, and translation evaluation. In the initial stage, parallel corpora are processed using the Moses Decoder software through a series of preprocessing steps. The two corpora are then cleaned, which is the cleaning process. They are then broken down into tokens, which is the tokenization process, and converted to all lowercase letters in the lowercase process. Moses Decoder is a machine translation application that implements SMT methods by utilizing probabilities in the language translation process, based on the analysis of parallel corpora [18]. During the training phase, both parallel and monolingual corpora are processed to generate language and translation models. The language model

serves as a textual knowledge foundation with probability values used to comprehend language structures and patterns. By studying a vast amount of text from both corpora, this model can identify probable words or phrases that may appear based on their context.

The core component of the language model is the probability of word sequences, expressed as a statistical value reflecting the likelihood of specific word sequences occurring in a language. This probability illustrates how often the word sequence appears in the text corpus used as the basis for learning the language model [19]. By understanding the probability of word sequences, the translation system can produce more accurate results that align with the language structure being used. The language model assigns probabilities to sequences of n words in a specific order with a uniform probability distribution. These sequences can be phrases or sentences [9]. One common approach to language modeling is the n-gram model. The n-gram language model is a type of probabilistic language model used to predict the next item in a sequence by considering a number of n-1 preceding items. Conditional probabilities can be computed using the frequency of n-grams occurring in the text corpus [20]. The frequency of n-grams refers to how often a specific sequence of words appears in the text data. Therefore, the more frequently an n-gram appears in the text corpus, the higher the likelihood that the sequence will appear in the next sentence.

During the training process, the translation model is also developed to help the system understand the relationship between the source language and the target language. In SMT, there are two translation models: word-based translation model and phrase-based translation model[21].

The next step is the decoding stage. In this stage, the system is tasked with finding the text or sentence in the target language with the highest probability, considering factors from the translation model and the previously trained language model. The source language, or Indonesian, will be processed, and the machine will generate translations into the target language, which is Bengkulu Malay. In the final step, the evaluation results are conducted using the Bilingual Evaluation Understudy (BLEU) software. This evaluation involves the translations generated in the testing phase, which are automatically evaluated to assess the quality of the machine translations in this research.

### 2.3 Testing and Evaluation

The testing of the implementation results of SMT from Indonesian to Bengkulu Malay was conducted using five different testing scenarios. The testing and evaluation processes were carried out automatically using the BLEU method. This evaluation helps assess how well the quality of machine translation achieves the desired objectives.

### 2.4 Analysis of Test Result

The analysis of the testing results in this study is used to examine the impact of parallel corpus usage on the quality and evaluation scores of SMT. Considering the quantity of parallel corpora used, this research aims to identify improvements in translation quality and evaluate whether the utilization of larger parallel corpora leads to significant enhancements in SMT performance.

## 3. RESULT AND DISSCUSION

In this section, describe in detail the research that has been conducted, discussing various sub-topics that encompass different stages of the research, such as data preprocessing, modeling processes, model training, and result evaluation. Additionally, we will present some of the findings and analyses of this research in the form of images and tables, to provide a clearer visual representation and support for the results we have obtained. Therefore, the following are the results and discussions generated from this research, expected to contribute significantly to the understanding and development in this field.

### 3.1. Implementation Statistical Machine Translation Indonesian to Bengkulu Malay

The first step in developing a statistical machine translator from Indonesian to Bengkulu Malay is corpus pre-processing. Pre-processing steps of the corpus that must be undertaken in the statistical machine translation process from Indonesian to Bengkulu Malay. This stage is crucial for preparing raw data before further processing in the translation system. This process involves several steps, including cleaning, lowercase conversion, and tokenization [22][23]. Through this pre-processing stage, it is expected that the corpus used in statistical machine translation can be more prepared and clean, thus resulting in more accurate and consistent translations [24].

Cleaning is an essential step in data pre-processing, carried out to remove empty sentences or reduce excessive space characters. Additionally, the cleaning process aims to enhance data quality by identifying and addressing anomalies, such as removing overly long sentences, which can hinder training model efficiency [25][26]. Therefore, optimizing the cleaning stage can contribute positively to the final results of the developed model [27]. The cleaning command can be viewed in the figure 3.

```
~/Desktop/smt/moses/scripts/training/clean-corpus-n.perl corpus/corpus id bkl corpus/clean 1 5000
```

Figure 3. Cleaning Command

The implementation of lowercase conversion is a commonly used method in text processing. The purpose of this process is to streamline the text by converting all letters to lowercase without altering the structure of the words [28]. The use of lowercase also helps reduce ambiguity in text processing, thus facilitating the development of more effective models or algorithms in tasks such as natural language processing and machine learning [29]. Lowercase contributes to improving the quality and consistency of data in the context of text analysis and language modeling [30]. The lowercase command can be viewed in the figure 4.

```
~/Desktop/smt/moses/scripts/tokenizer/lowercase.perl < corpus/clean.id > corpus/lowercased.id
~/Desktop/smt/moses/scripts/tokenizer/lowercase.perl < corpus/clean.bkl > corpus/lowercased.bkl
```

Figure 4. Lowercase Command

In general, tokenization involves the process of breaking a sequence of characters into word units to create a more structured representation. This stage not only separates words but also plays a role in cleaning the text by removing punctuation, numbers, and other characters that are not part of the alphabet [31][32]. Thus, each word can stand alone as a separate entity. The concept of tokenization involves the recognition and grouping of linguistic elements in the text into smaller units called tokens. Tokens can include punctuation marks, words, numbers, or other elements that have meaning within the context of the text [33]. Tokenization process is often employed to prepare text data before it is passed through natural language processing models or algorithms. By utilizing tokenization techniques, text analysis becomes more manageable as the text has been transformed into a sequence of structured tokens [34]. The tokenization command can be viewed in the figure 5.

```
~/Desktop/smt/moses/scripts/tokenizer/tokenizer.perl < corpus/lowercased.id > corpus/tokenized.id
~/Desktop/smt/moses/scripts/tokenizer/tokenizer.perl < corpus/lowercased.bkl > corpus/tokenized.bkl
```

Figure 5. Tokenization Command

After completing all pre-processing stages, the next step is to enter the training phase. In this stage, the construction of language and translation models is conducted. The language model training process aims to achieve the target language model, specifically in the context of Bengkulu Malay. The language model is executed using the SRI Language Modeling (SRILM) software, which has been integrated with the Moses Decoder [35]. In this stage, the language model will generate n-gram values and calculate their probabilities, forming a crucial foundation for understanding and generating text in the targeted language. The languange model command can be viewed in the figure 6.

```
~/Desktop/smt/srilm/bin/i686-m64/ngram-count -order 3 interpolate -unk -text corpus/tokenized.bkl -
lm lm/bkl.lm
```

Figure 6. Language Model Command

From the trained language model, this process will generate a file with the extension ".lm" and save it as "output.bkl.lm". The language model creates n-gram data consisting of unigrams (n-gram 1), bigrams (n-gram 2), and trigrams (n-gram 3). Unigram contains one token, bigram contains two tokens, and trigram contains three tokens. Additionally, each n-gram data also includes its probability values. Figure 7 below is an example of the language model produced by SRILM on the statistical machine translator for Indonesian to Bengkulu Malay language.

```
\data\
ngram 1=3276
ngram 2=16256
ngram 3=2130

\1-grams:
-2.046147       apo     -0.4858913
-3.521818       bahayo  -0.1438879
-3.521818       cuaco   -0.08475756
-----------------------------------------------------------------------------
---
\2-grams:
-1.636891       beli rubo-rubo      -0.01463723
-0.845098       cakmano caro        -0.04331619
-1.01424        dimano ado          -0.3303767
-----------------------------------------------------------------------------
----
\3-grams:
-0.2777181      kecek mak ambo
-0.6858329      ambo namo nyo
-0.5608942      biso nyari caro
```

Figure 7. SRILM Generated Language Model

The next step is the translation model stage using the tool available in Moses, namely Giza++. The data utilized is the parallel corpus between Indonesian and Bengkulu Malay. The translation model process produces vocabulary, word alignment adjustment, and translation model. The translation model command can be viewed in the figure 8.

```
~/Desktop/smt/moses/scripts/training/train-model.perl -root-dir . --corpus corpus/tokenized --f id --e
bkl --lm 0:3:$PWD/lm/bkl.lm -external-bin-dir ~/Desktop/smt/training-tools
```

Figure 8. Translation Model Command

In the translation model process, GIZA++ generates vocabulary. Numbers 1 to 20 in the corpus vocabulary are unique IDs for each data token, while the numbers to the right of the tokens indicate how often the word appears. The function of the vocabulary generated by GIZA++ is to show words with the highest frequency. For example, in the vocabulary illustration above, the word "kau" has the highest frequency, which is 1122. Table 2 is an example of vocabulary.

Table 2. An example of vocabulary

| Uniq id | Word | Frequency |
|---|---|---|
| 1 | UNK | 0 |
| 2 | ambo | 2063 |
| 3 | kau | 1122 |
| 4 | ? | 900 |
| 5 | iko | 660 |
| 6 | idak | 575 |
| 7 | yang | 487 |
| 8 | nian | 466 |

| | | |
|---|---|---|
| 9 | nyo | 395 |
| 10 | di | 317 |
| 11 | apo | 299 |
| 12 | kek | 292 |
| 13 | ado | 287 |
| 14 | banyak | 267 |
| 15 | ndak | 246 |
| 16 | ko | 223 |
| 17 | kito | 218 |
| 18 | pai | 213 |
| 19 | . | 193 |
| 20 | rumah | 185 |

In addition to the vocabulary, there is a word alignment document to illustrate the mapping of words from the source sentence to the target sentence. The word alignment document illustrates the mapping of words from the source sentence to the target sentence. In the Indonesian-Bengkulu Malay word alignment, there are three lines of sentences. The first line provides information about the position of the target sentence (3) in the corpus, the length of the source sentence (6), the length of the target sentence (6), and the alignment value (0.0660107). The figure 9 below is an example of word alignment in Bengkulu Malay Language.

```
# Sentence pair (37) source length 6 target length 6 alignment score : 0.0660107
ambo datang megundang kau anak beranak
NULL ({ }) ambo ({ 1 }) datang ({ 2 }) megundang ({ 3 }) kau ({ 4 }) anak ({ 5 }) beranak ({ 6 })
……
```

Figure 9. Word Aligment Bengkulu Malay Language

The phrase-based translation model table plays a role in aligning input text from Indonesian to output text in Bengkulu Malay. The translation process in the phrase-based translation model can be broken down into several stages, including dividing the source language sentences into phrases, translating each phrase into the target language, and recording or storing the resulting sentences. In the process of segmenting sentences into phrases, the source language will be aligned with words in the target language. Example Pieces of the Indonesian-Bengkulu Malay Translation Model can be seen in figure 10 below.

```
ada halangan hari raya nanti kami ingin ||| ado halangan rayo-rayo kelak kami ndak ||| 1 0.146 907 1
0.161493 ||| 0-0 1-1 2-2 3-2 4-3 5-4 6-5 |||
1 1 1 ||| |||
ada hutang dengan kamu seingat saya ||| ado utang kek kau seingek  ambo ||| 1 0.783433 1|||  0.594015
||| 0-0 1-1 2-2 3-3 4-4 5-5 ||| 1 1 1 |||
 , saya membeli ini dan ternyata rusak ||| , ambo ngebeli iko dan ternyato rusak ||| 1 0.874151 1 0.192577
||| 0 -0 1-1 2-2 3-3 4-4 5-5 6-6 ||| 1 1 1 ||| |||
, tapi saya melakukannya di sma ||| , tapi ambo ngelakukannyo di sma  ||| 1 0.84512 1 0.79946 ||| 0-0 1-
1 2-2 3-3 4-4 5-5 ||| 1 1 1 ||| ||| , tol ong beri saya pesanan baru . ||| , tolong kasih ambo pesanan baru .
||| 1 0.314706 1 0.702736 ||| 0-0 1-1 2-2 3-3 4-4 5-5 6-6 ||| 1 1 1 ||| |||
```

Figure 10. Pieces of the Indonesian-Bengkulu Malay Translation Model

In the decoding process, input sentences from the source language, which is Indonesian, will undergo the translation process. Subsequently, these input sentences will be processed using the Moses Decoder. The result is sentences in the target language, which is Bengkulu Malay. The Moses Decoder plays a crucial role in converting text from the source language to the target language. The decoding command can be viewed in the Figure 11.

```
~/Desktop/smt/moses/bin/moses -f model/moses.ini
```

Figure 11. Decoding Command

The image below shows some examples of translation tests on the Moses decoder, starting with the first sentence. The phrase "nenek saya masak nasi hari ini" is translated into "nenek ambo masak nasi ari iko". The Moses Decoder will search for the translation output from the source sentence considering the statistical language model and translation model with the highest probabilities. The language model's role is to ensure fluency in translation, while the translation model's role is to ensure accuracy in the translation process. The example translation using moses decoder can be seen in Figure 12.

```
nenek saya masak nasi hari ini
Translating: nenek saya masak nasi hari ini
Line 1: Initialize search took 0.000 seconds total
Line 1: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 1: Search took 0.034 seconds
nenek ambo masak nasi ari ko
BEST TRANSLATION: nenek ambo masak nasi ari ko [111111]  [total=-7.804] core=(0.
000,-6.000,6.000,-0.059,-0.132,-0.900,-1.045,0.000,-29.155)
Line 1: Decision rule took 0.000 seconds total
Line 1: Additional reporting took 0.001 seconds total
Line 1: Translation took 0.035 seconds total
```

Figure 12. Example Translation Using Moses Decoder

### 3.2. Evaluating

The system evaluation is conducted through automatic testing on SMT, which produces accuracy using the BLEU method. BLEU is an automatic evaluation metric that works by comparing the output results of the text machine translation system (candidates) with translations obtained from humans (references) [36][37]. The main task of BLEU is to compare candidate n-grams with reference n-grams and calculate the number of matching words [38]. The more words appear, the better the translation. The formula is:

$$\text{BP}_{\text{BLEU}} = f(x) = \begin{cases} 1, & if\ c > r \\ e^{(1-\frac{r}{c})}, & if\ c \leq r \end{cases} \quad (1)$$

$$Pn = \frac{\sum C \in \{Candidates\} \sum ngram \in C countclip(ngram)}{\sum C \in \{Candidates\} \sum ngram \in C countclip(ngram)} \quad (2)$$

$$\text{BLUE} = \text{BP}.\exp \sum_{n=1}^{N} Wn.logPn \quad (3)$$

Information:
BP = brevity penalty
C = number of words from automatic translation
R = number of references
Pn = Modified precision score
Wn = 1/N (N for BLUE which is 4)
Pm = The number of n-grams translated by reference divided by the number of n-grams translated.

The testing on this system is evaluated based on how closely the translations produced by the Moses decoder approximate the translations considered correct according to the reference. This evaluation helps assess the quality and performance of the machine translation system in producing accurate translations that meet user needs.

This research employs five scenarios. In the first scenario, 500 sentences from the Indonesian - Bengkulu Malay parallel corpus are used to test the translator system's performance. The second scenario extends the testing by involving 1500 sentences from the same corpus, while the third scenario involves 2500 parallel corpus. In the fourth scenario, 4000 sentences from the corpus are used with the expectation that additional data will provide a better understanding of the language

structure and context. Meanwhile, the fifth scenario takes a step further by using 5261 sentences from the same parallel corpus, aiming for a significant improvement in translator system accuracy and reliability. The results of these five scenarios will be elaborated in detail in table 3 below, providing a deeper understanding of how the parallel corpus size affects the translator system's performance.

Table 3. Measuring BLEU Scores for Different Scenarios

| Machine Translation | Parallel Corpus Quantities | BLEU Score |
|---|---|---|
| Scenario 1 | 500 | 80,56 % |
| Scenario 2 | 1500 | 90, 86 % |
| Scenario 3 | 2500 | 92,50 % |
| Scenario 4 | 4000 | 92,91 % |
| Scenario 5 | 5261 | 94, 48 % |

From the results seen in Table 3, it can be concluded that increasing the number of parallel corpora in each scenario, namely the first, second, and third scenarios, provides the potential for significant improvement in translation accuracy from Indonesian to Bengkulu Malay. This change indicates that the more data available for training, the better the translator system can understand and interpret text accurately. This suggests that the use of larger parallel corpora can be one effective strategy in improving the quality of inter-language translation.

The experiments were conducted in five different scenarios to evaluate the performance of the translator system. The results from each scenario show different levels of accuracy, with the first scenario achieving , the second scenario achieving , the third scenario achieving , the fourth scenario achieving , and the fifth scenario achieving a high BLEU score of . This BLEU score is a common metric used to measure the quality of machine translation by comparing machine translations with reference translations made by humans. Although the third scenario showed a significant improvement in accuracy, further details regarding the results of this scenario will be presented through examples listed in table 4 below.

Table 4. Example Translation of Fifth Scenario

| Indonesian Input Sentences | Reference Senteces in Bengkulu Malay Languange | Result from SMT Bengkulu Malay Languange |
|---|---|---|
| Saya kira hari minggu | Ambo kiro ari minggu | Ambo kiro hari minggu |
| Saya sangat lapar sekarang | Ambo lapar nian kini | Ambo lapar nian kini |
| Cepat sekali kamu ini berjalan nya | Cepek nian kau iko bejalan nyo | Cepek nian kau ko bejalan nyo |
| Dirumah ini saya akan membuat acara | Dirumah iko ambo akan ngebuek acaro | Dirumah ko ambo akan ngebuek acaro |
| Kamu ingin pergi kemana siang hari ini | Kau ndak pai kemano siang ari iko | Kau ndak pai kemano siang hari ko |

The data found in Table 4 shows five sample test sentences used in this study from the fifth scenario. The first column presents the test sentences in Indonesian, while the second column contains reference sentences for the translation of the test sentences into Bengkulu Malay, and the third column displays the translation results generated by SMT. In the example of the first sentence from the fifth scenario, there appears to be a slight difference between the translation result and the expected reference. In the first sentence, SMT translates the word "hari" in Indonesian as "hari," whereas it should be translated as "ari" in Bengkulu Malay. In the second sentence example, SMT produces a translation that matches the reference. However, in testing the third, fourth, and fifth sentences, SMT is less accurate in translating the input word "ini," which should be translated as "iko," but SMT provides "ko" as the result.

## 4. CONCLUSION

The results of this research indicate that the amount of parallel corpus usage significantly influences the development of machine translation and contributes to improving the accuracy of translation results. However, the availability of parallel corpora for Indonesian and Bengkulu Malay

language pairs is much more limited compared to other language pairs that have larger and more reliable resources. The use of 500 parallel corpora resulted in a BLEU score of 80.56%. When the number of parallel corpora was expanded to 1500, the BLEU score increased to 90.86%, and with 2500 corpora, the BLEU score reached 92.50%. With 4000 parallel corpora, the BLEU score increased to 92.91%, indicating a further improvement in translation accuracy. However, increasing the quantity of parallel corpora to 5261 resulted in a higher BLEU score of 94.48%, indicating that the larger the number of parallel corpora used, the better the performance of the translation system.

## ACKNOWLEDGEMENTS

This research was supported by the Direktorat Riset, Teknologi, dan Pengabdian Masyarakat (DRTPM) Direktorat Jenderal Pendidikan Tinggi, Riset, dan dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi under Penelitian Tesis Magister (Master's Thesis Research) with grant number: 061/PTM/LPPM-UAD/VI/2024 (15 Juni 2024).

## REFERENCES

[1] J. Zakaria, I. Yuniati, and E. F. Wijaya, "Implikatur Tegur Sapa Dalam Bahasa Melayu Bengkulu," *Lit. J. Bahasa, Sastra dn Pengajaran*, vol. 1, no. 2, pp. 74–78, 2021, doi: https://doi.org/10.31539/literatur.v1i2.2401.

[2] Asrif, "Pembinaan dan Pengembangan Bahasa Daerah dalam Memantapkan Kedudukan dan Fungsi Bahasa," pp. 11–23, 1945.

[3] N. H. M. Ningsih, D. E. C. Wardhana, and S. Supadi, "Derivasi Bahasa Melayu Bengkulu," *J. Ilm. KORPUS*, vol. 4, no. 2, pp. 224–230, 2020, doi: 10.33369/jik.v4i2.8361.

[4] F. Senovil, "Morfofonemik Bahasa Melayu Bengkulu," *KLITIKA J. Ilm. Pendidik. Bhs. dan Sastra Indones.*, vol. 2, no. 2, pp. 165–178, 2020, doi: https://doi.org/10.32585/klitika.v2i2.1037.

[5] R. Afria, J. Izar, R. D. Anggraini, and D. H. Fitri, "Analisis Komparatif Bahasa Bengkulu, Rejang, Dan Enggano," *Ling. Fr. Bahasa, Sastra, dan Pengajarannya*, vol. 5, no. 1, p. 1, 2021, doi: 10.30651/lf.v5i1.4274.

[6] D. E. C. Wardhana, D. Kusumaningsih, and A. C. S. Dewi, " Model of Perception and Critical Language Style of Academic Community at University of Bengkulu During Coronavirus Disease (COVID) 19 Epidemic to Realize the Freedom of Learning ," vol. 485, no. Iclle, pp. 223–227, 2020, doi: 10.2991/assehr.k.201109.038.

[7] A. Sudarsono, "Jaringan Syaraf Tiruan Untuk Memprediksi Laju Pertumbuhan Penduduk Menggunakan Metode Bacpropagation (Studi Kasus Di Kota Bengkulu)," *J. Media Infotama*, vol. 12, no. 1, pp. 61–69, 2016, doi: 10.37676/jmi.v12i1.273.

[8] E. Widianto, "Pemertahanan Bahasa Daerah melalui Pembelajaran dan Kegiatan di Sekolah," *J. Kredo*, vol. (1) 2, pp. 1–13, 2018.

[9] R. Darwis, H. Sujaini, and R. D. Nyoto, "Peningkatan Mesin Penerjemah Statistik dengan Menambah Kuantitas Korpus Monolingual ( Studi Kasus : Bahasa Indonesia – Sunda )," vol. 7, no. 1, pp. 27–32, 2019.

[10] A. E. P. Lesatari, A. Ardiyanti, and I. Asror, "Phrase Based Statistical Machine Translation Javanese-Indonesian," *J. Media Inform. Budidarma*, vol. 5, no. 2, pp. 378–386, 2021, doi: http://dx.doi.org/10.30865/mib.v5i2.2812.

[11] Permata and Z. Abidin, "Statistical Machine Translation Pada Bahasa Lampung Dialek Api Ke Bahasa Indonesia," *J. Media Inform. Budidarma*, vol. 4, no. 3, pp. 519–528, 2020, doi: http://dx.doi.org/10.30865/mib.v4i3.2116.

[12] Q. A. Agigi and A. A. Suryani, "Statistical Machine Translation Muna to Indonesia Language," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 4, pp. 2173–2186, 2021, doi: 10.35957/jatisi.v8i4.1149.

[13] M. S. Alam and A. A. Suryani, "Minang and Indonesian Phrase-Based Statistical Machine Translation," *J. Informatics Telecommun. Eng.*, vol. 5, no. 1, pp. 216–224, 2021, doi: https://doi.org/10.31289/jite.v5i1.5308.

[14] M. F. Khaikal and A. A. Suryani, "Statistical Machine Translation Dayak Language – Indonesia Language," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 16, no. 1, pp. 49–56, 2021, doi: http://dx.doi.org/10.30872/jim.v16i1.5315.

[15] S. M. A. Razak, M. S. A. Seman, W. Ali, W. Y. Wan, N. H. Nizan, and M. Noor, "Malay manuscripts transliteration using statistical machine translation (SMT)," *Proc. - 2019 1st Int. Conf. Artif. Intell. Data Sci. AiDAS 2019*, pp. 137–141, 2019, doi: 10.1109/AiDAS47888.2019.8970867.

[16] A. Jannesari, "Statistical Machine Translation Outperforms Neural Machine Translation in Software Engineering : Why and How," pp. 3–12, 2020, doi: 10.1145/3416506.3423576.

[17] N. S. Khan, A. Abid, and K. Abid, "A Novel Natural Language Processing (NLP)–Based Machine Translation Model for English to Pakistan Sign Language Translation," *Cognit. Comput.*, vol. 12, no. 4, pp. 748–765, 2020, doi: 10.1007/s12559-020-09731-7.

[18] M. N. Amin, A. B. P. Negara, and A. Perwitasari, "Implementasi Mesin Penerjemah Statistik Pada Aplikasi Chatting Berbasis AndroidDengan Moses Decoder," *Infotekjar J. Nas. Inform. dan Teknol. Jar.*, vol. 6, no. 1, pp. 155–164, 2021, [Online]. Available: https://jurnal.uisu.ac.id/index.php/infotekjar/article/view/4025/0.

[19] A. M. Gezmu, A. Nürnberger, and T. B. Bati, "Extended Parallel Corpus for Amharic-English Machine Translation," *2022 Lang. Resour. Eval. Conf. Lr. 2022*, pp. 6644–6653, 2022.

[20] J. Liu, "Comparing and Analyzing Cohesive Devices of SMT and NMT from Chinese to English: A Diachronic Approach," *Open J. Mod. Linguist.*, vol. 10, no. 06, pp. 765–772, 2020, doi: 10.4236/ojml.2020.106046.

[21] M. Wahyuni, H. Sujaini, and H. Muhardi, "Pengaruh Kuantitas Korpus Monolingual Terhadap Akurasi Mesin Penerjemah Statistik," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 1, pp. 20–26, 2019, doi: https://dx.doi.org/10.26418/justin.v7i1.27241.

[22] Z. Yu, Z. Yu, J. Guo, Y. Huang, and Y. Wen, "Efficient Low-Resource Neural Machine Translation with," vol. 19, no. 3, pp. 1–13, 2020.

[23] H. Yuliansyah, S. A. Mulasari, S. Sulistyawati, F. A. Ghozali, and B. Sudarsono, "Sentiment Analysis of the Waste Problem based on YouTube comments using VADER and Deep Translator," *J. Media Inform. Budidarma*, vol. 8, pp. 663–673, 2024, doi: 10.30865/mib.v8i1.6918.

[24] M. Popel *et al.*, "Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals," *Nat. Commun.*, vol. 11, no. 1, pp. 1–15, 2020, doi: 10.1038/s41467-020-18073-9.

[25] H. Yuliansyah, M. Iqbal, and A. Latiffi, "Sentiment Analysis of the Sheikh Zayed Grand Mosque ' s Visitor Reviews on Google Maps Using the VADER Method," vol. 5, no. 1, 2024, doi: 10.59395/ijadis.v5i1.1320.

[26] F. Rahutomo, A. A. Septarina, M. Sarosa, A. Setiawan, and M. M. Huda, "A review on Indonesian machine translation," *J. Phys. Conf. Ser.*, vol. 1402, no. 7, 2019, doi: 10.1088/1742-6596/1402/7/077040.

[27] A. Bandyopadhyay, I. Kundu, A. Chakraborty, R. Kumar, A. Kumar, and S. Sabut, *Blood Donation Management System Using Android Application*, vol. 728 LNEE. 2021.

[28] I. Factor, "RNN Encoder or Decoder-Based Phrase Representation Learning For," no. 1, pp. 1–10, 2023.

[29] R. Achmad, Y. Tokoro, J. Haurissa, and A. Wijanarko, "Recurrent Neural Network-Gated Recurrent Unit for Indonesia-Sentani Papua Machine Translation," *J. Inf. Syst. Informatics*, vol. 5, no. 4, pp. 1449–1460, 2023, doi: 10.51519/journalisi.v5i4.597.

[30] T. I. Ramadhan, N. G. Ramadhan, and A. Supriatman, "Implementation of Neural Machine Translation for English-Sundanese Language using Long Short Term Memory (LSTM)," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1438–1446, 2022, doi: 10.47065/bits.v4i3.2614.

[31] D. W. Otter, J. R. Medina, and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 2, pp. 604–624, 2021, doi: 10.1109/TNNLS.2020.2979670.

[32] Y. Dong, "RNN Neural Network Model for Chinese-Korean Translation Learning," *Secur. Commun. Networks*, vol. 2022, 2022, doi: 10.1155/2022/6848847.

[33] J. Xiao and Z. Zhou, "Research Progress of RNN Language Model," *Proc. 2020 IEEE Int. Conf. Artif. Intell. Comput. Appl. ICAICA 2020*, pp. 1285–1288, 2020, doi: 10.1109/ICAICA50127.2020.9182390.

[34] A. Othman and M. Jemni, "Designing high accuracy statistical machine translation for sign language using parallel corpus: Case study English and American Sign language," *J. Inf. Technol. Res.*, vol. 12, no. 2, pp. 134–158, 2019, doi: 10.4018/JITR.2019040108.

[35] Z. Abidin, "Penerapan Neural Machine Translation untuk Eksperimen Penerjemahan secara Otomatis pada Bahasa Lampung – Indonesia," *Pros. Semin. Nas. Metod. Kuantitatif*, no. 978, pp. 53–68, 2017.

[36] M. Gerdy Asparilla, H. Sujaini, R. Dwi Nyoto, and J. H. Hadari Nawawi, "Perbaikan Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik (Studi Kasus : Bahasa Indonesia-Jawa Krama)," vol. 1, no. 2, 2018.

[37] Y. Jarob, H. Sujaini, and N. Safriadi, "Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman Dengan Penandaan Kata Dasar Dan Imbuhan," *J. Edukasi dan Penelit. Inform.*, vol. 2, no. 2, pp. 78–83, 2016, doi: 10.26418/jp.v2i2.16520.

[38] H. Sujaini, "Peningkatan Akurasi Penerjemah Bahasa Daerah dengan Optimasi Korpus Paralel," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 1, 2018, doi: 10.22146/jnteti.v7i1.394.