

Research Article

The Application of Computer-Aided Under-Resourced Language Translation for Malay into Kadazandusun

Mohd Shamrie Sainin^{1,*}, Minah Sintian², Suraya Alias¹ and Asni Tahir¹

¹Language Engineering and Application Development, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia

shamrie@ums.edu.my; suealias@ums.edu.my; asnieta@ums.edu.my

²Faculty of Languages and Communication, Universiti Pendidikan Sultan Idris, Malaysia

minahsintian@fbk.upsi.edu.my

*Correspondence: shamrie@ums.edu.my

Received: 1st September 2022; Accepted: 5th May 2023; Published: 5th October 2023

Abstract: A computer-aided language translation using a Machine translation (MT) is an application performed by computers (machines) that translates one natural language to another. There are many online language translation tools, but thus far none offers a sequence of text translations for the under-resourced Kadazandusun language. Although there are web-based and mobile applications of Kadazandusun dictionaries available, the systems do not translate more than one word. Hence, this paper aims to present the discussion of the preliminary translation of Malay to Kadazandusun. The basic word-to-word with dictionary alignment translation based on Direct Machine Translation (DMT) is selected to begin the exploration of the translation domain where DMT is one of the earliest translation methods which relies on the word-to-word approach (sequence-to-sequence model). This paper aims to investigate the under-resourced language and the task of translating from the Malay language to the Kadazandusun language or vice versa. This paper presents the application and the process as well as the results of the system according to the basic Kadazandusun word arrangement (Verb-Subject-Object) and its translation quality using the Bilingual Evaluation Understudy (BLEU) score. Several phases are involved during the process, including data collection (word pair translation), preprocessing, text selection, translation procedures, and performance evaluation. The preliminary language translation approach is proven to be capable of producing up to 0.5 BLEU scores which indicate that the translation is readable, however, requires post-editing for better comprehension. The findings are significant for the quality of the under-resourced language translation and as a starting point for other machine translation methodologies such as statistical or deep learning-based translation.

Keywords: *Computer-aided; Kadazandusun; Language Translation; Machine Translation; Malay; Under-resourced*

1. Introduction

There are many existing web-based MT systems available online. However, based on analysis gathered during the observation, although some of the translation services provide translation in the Malay language, none of them offer text translation in the Kadazandusun language which is one of the native languages in Sabah and classified as an under-resourced language. The language is currently taught in several schools and public universities, and it is used in a local newspaper's news section. Translation of a low-resource language like Kadazandusun is limited to dictionary lookup for one word and not for a sentence or a certain length of text in a single translation request. As a result, the major purpose of this research is to examine preliminary language translation utilizing the DMT technique. Machine translation (MT) refers to a computerized linguistic translation from one language to another (from a source language or SL and target language or TL) [1], where it is a subfield of Artificial

Intelligence for automatic language translation. It is also a multidisciplinary domain of research and application with support from different schools of thought such as computer science, artificial intelligence, mathematical modeling, statistics, education, language, linguistics, and many more.

The goal of machine translation (MT) is to employ software or technology to allow individuals to convert a written language to another (textual pair). In Malaysia, the Malay language has remained a core language, with the emphasis on developing improved MT. The earliest publication related to a translation system for the Malay-English language can be found in [2]. Furthermore, several papers specifically addressing the research in MT for the Malay language were slowly gaining attention as summarized in Table 1.

Table 1. Chronology of language translation centered on the Malay language

Reference	Translation	Description/Domain
Cheong (1986) [2]	English-Malay	The first project on a computer-aided translation system for Malay-English started at Universiti Sains Malaysia, based on a secondary school chemistry textbook.
Cheong (1987) [3]	English-Malay	A study on the interrogative model in a Malay-English translation system.
Ogura <i>et al.</i> (1999) [4]	Japanese-Malay	Semantic-based transfer MT system applied to a Japanese to Malay translation prototype using 20,000 Malay entries.
Yeong <i>et al.</i> (2016) [1]	English-Malay	Statistical MT, applying a dictionary and lemmatizer.
Wang <i>et al.</i> (2016) [5]	Malay/Indonesian- English	Improved statistical-based MT for resource-poor MT.
Alsaket and Aziz (2014) [6]	Arabic-Malay	A rule-based approach was applied to develop the MT system with quite good human judgment accuracy at 92.3%.
Almeshrky and Aziz (2012) [7]	Arabic-Malay	A transfer-based approach which implemented to the MT system with 89.4% accuracy, comparing human judgment and system translation.
Lakew <i>et al.</i> (2018) [8]	Varieties/ Indonesian-Malay	A neural MT approach study on varieties of language including Indonesian-Malay.
Chua <i>et al.</i> (2018) [9]	English-Malay	Example-based MT combined with analogical-based and structural semantics in English-Malay translation. The reported BLEU score is about 37.06%.

MT is a system associated with converting a text from one language to another with similar or comparable meaning and grammatical structure. In comparison to Malay language translation, the Kadazandusun language currently has a limited state-of-the-art MT study. For the time being, researchers in this field have undertaken preliminary studies in order to create a system that would leverage existing methodologies and techniques to discover a viable translation methodology. Resources available online and offline that use the Kadazandusun language could not benefit the community because the population is not able to fully understand the language, especially the younger generation. With about 30 percent of the entire population in Sabah, the Kadazandusun are the biggest ethnic community. In response to this issue, this paper proposes using rule-based direct translation to study MT from Kadazandusun language texts to Malay or vice versa. The purpose of this research is to assess the source language utilizing text from local newspapers (Malay articles) with translation into Kadazandusun as training data.

Direct Machine Translation, Rules, Corpus, Statistical, and Transfer Approach are the five major methodologies in MT. From these approaches, the last four require a parallel text corpus to generate their model (e.g., rules, analogy, and statistics), while the Direct Machine Translation (DMT) approach is a bilingual and uni-directional from the source to the target language. DMT is a word-by-word sequence translation method that may incorporate structural or grammatical changes. Although DMT is less capable of translating sequences of text effectively, this study investigates the preliminary MT work using this method for under-resourced languages like the Kadazandusun. Furthermore, this is known to be the first attempt to use computing capabilities to apply MT from Bahasa Melayu to Kadazandusun or vice versa for a sentence translation. Thus, the basic MT approach using word-to-word translation is investigated.

Various approaches used in MT have been classified into two: the single approach and the hybrid approach. The single approach is defined as employing only one way in the translation, while the hybrid

incorporates the statistical method with a rule-based approach that includes several frameworks such as syntax, forest, word, and phrase [10].

DMT is regarded as a fundamental method for translating a sequence of words from one language to another without much linguistic processing, utilizing a bilingual dictionary. It is also known as a dictionary-driven MT. The translation process for DMT is depicted in Figure 1. In the DMT system, the morphological analysis will extract all words from the text in a source language. It may involve preprocessing (removing unwanted characters, etc.) and root word generation with stemming. The second step is to look up a base word or an original word in the pairwise bidirectional dictionary. The dictionary must have a match of pairwise words for words in the target language or unsuccessful translation otherwise. The final step is to perform some degree of syntactic rearrangement of the words according to the predefined rules in the system. This phase will rearrange the TL words to match the sentence in the target language and output it as a TL text.

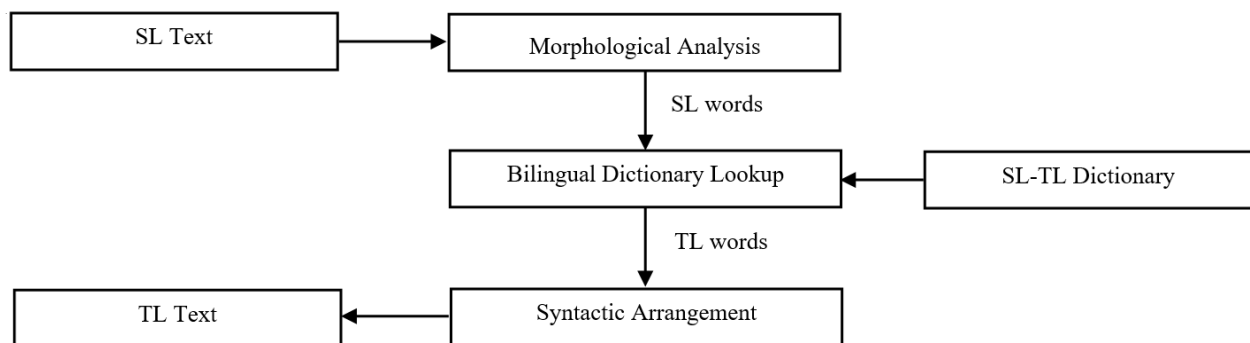


Figure 1. DMT System [11]

The quality of MT can be measured using evaluation metrics such as BLEU [12], NIST [13], METEOR [14], Perplexity Matrix [15], and Neural Network [16]. Among these evaluation metrics, BLEU (Bilingual Evaluation Understudy) is a term that appears often in MT literature. The scoring value using BLEU is a method for evaluating MT automatically using these features: quick and inexpensive calculation, easy to understand, language-independent, high correlation with human evaluation, and widely adopted. The following Equation 1 is used to determine the BLEU score.

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

$$BLEU = BP \cdot \exp(\sum_{n=1}^N w_n \log p_n)$$

In Equation 1, the application of Brevity Penalty (BP) in BLEU, where r is the words total from reference, q is the words total in the translated output, N is the value of n-grams (1-gram, bigram, 3-gram, 4-gram), w_n is the weight of the precision and p_n is the modified precision.

BLEU calculates the similarity between reference and created phrases using the fundamental notions of n-gram precision. The scale range used in the BLEU metric is from 0 to 1 and is normally presented in a percentage value (0%-100%). Closer to 1 indicates that the translation corresponds to a human translation. A BLEU score of less than 50% or 0.5, indicates that the MT engine is poor and not performing optimally, resulting in a higher level of post-editing required before reaching publishable quality. While a score of 50% to 100%, or (0.5-1.0), is generally considered an average translation with some post-editing required [12].

2. Material and Methods

As stated in the introduction, there are currently no MT systems available for the Kadazandusun language. As a result, this is the first attempt to use the DMT approach to investigate an MT in Kadazandusun. Any MT would include the following steps: 1) source language text input, 2) source language text decoding (also known as the transfer phase), and 3) text translation or encoding to the target language. In step 1, most systems will do text preprocessing with the existence of a corpus or a word database. The next step is using certain approaches to decode the text such as pairwise or dictionary-based, sequence to sequence, and neural-based decoding. Finally, in step 3, the translation is simply encoding the source text to a target text. The decoder will create a translation structure that may preserve

the meaning of the original sentence to a target language, ranging from simple word-to-word translation with some linguistic alignments to complex neural decoding structures.

The preliminary study on MT for the Kadazandusun language is done using DMT word-to-word translation with dictionary and rule alignment. The steps involved in the procedure are depicted in Figure 2 below.

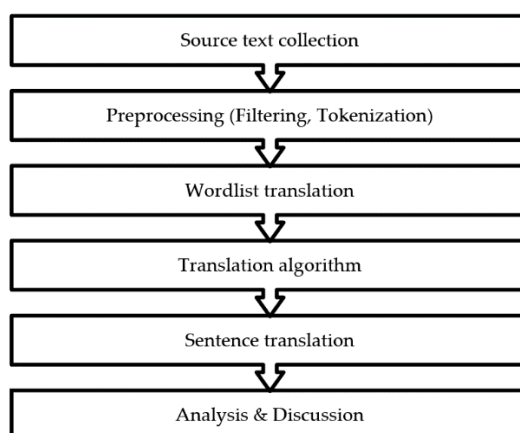


Figure 2. The process in the translation system investigation

2.1. Source Text Collection

The first step is the gathering of a text corpus, where the preliminary corpus is a Kadazandusun newspaper section that contains publicly available news acquired from the New Sabah Times. Table 2 provides an overview of the corpus information. Text preprocessing is applied to the text to get the Kadazandusun word list. The first stage of text preprocessing was applied based on the best settings observed in [17]. The filtering procedure removes characters that match with integers and alphanumeric, leaving only word tokens in the text. Next, the tokenization process is applied to extract a single word based on a whitespace tokenizer. It provides approximately 14531 unique words; however, it also contains several words in English, Malay, and names that needed to be removed. After those words were removed, a total of 7079 remaining words were recorded for the next phase.

Table 2. Source text collection

Attribute	Details
Source	Kadazandusun news archive
Date of the news article collection	1/1/2018 – 12/7/2019
Number of documents	591
Number of words (tokens)	186569
Average words	315
Number of sentences	13641
The average length of sentence	23
Number of unique terms	14531

2.2. Word Translation Database

In this step, the main references book for creating the wordlist were the Kadazandusun dictionary [18], Kamus Malay-Dusun-English [19], and [20]. Based on the previous step, the list of 7079 words was used to produce a word-to-word match between the Kadazandusun and the Malay language. Out of 7079 words, 5663 of the words were not found or spelled differently in the dictionary. Thus, this poses a big challenge for word-to-word matching in the DMT.

3. Result and Discussions

The language translation algorithm based on DMT is tested with specific 34 sentences and the sample translation are listed in Table 3. The Kadazandusun language sequence is in the form of Verb-Subject-Object (VSO) as compared to the Malay language which is Subject-Verb-Object (SVO)[21], and 34 sentences are adequate to cover these sequences in this preliminary study. More sentences will be included in the future study which focuses on advanced phrase patterns. The results in Table 3 will be

discussed in four basic phrase patterns which are noun + verb, noun + noun, noun + prepositional, and noun + adjective.

Table 3. Translation sample Source text collection

No.	Source (Malay Language)	Translated to Kadazandusun Language using DMT
1.	Menteri berkata.	Montiri minoboros.
2.	Saya menjemur padi	Oku poposidang parai.
3.	Nenek tersenyum.	Odu kongingis.
4.	Kami berlari di sekolah.	Yahai manangkus id sikul.
5.	Saya terlupa memadam api.	Oku nolihuan momisok tapui.
6.	Saya akan beli ikan untuk Nancy	Oku atantu' boli sada montok Nancy.
7.	Kami suruh dia lari.	Dahai sinuhu disio tangkus.
8.	Semua pelajar telah diberi uniform.	Oinsanai susumikul telah nonuan uniform.
9.	Semua pemimpin masyarakat dijemput hadir.	Oinsanan puru ginumuan alapon rumikot.
10.	Jimmy Palikat anak kampung.	Jimmy Palikat tanak kampung.
11.	Leha menangis.	Leha mihad.
12.	Lihan sangat kurus.	Lihan tomod agahui.
13.	Sekolah saya di Ranau.	Sikul oku id Ranau.
14.	Hilda guru.	Hilda mongingia'.
15.	Bapa saya guru ugama.	Tapa oku mongingia' tugama.
16.	Tiga orang guru sekolah.	Tolu tulun mongingia' sikul.
17.	Pokok kelapa.	Puun piasau.
18.	Ibu bapa.	Tama tapa.
19.	Lelaki perempuan.	Kusai tondu.
20.	Alice membaca buku.	Alice mambasa buuk.
21.	Halim tersenyum.	Halim kongingis.
22.	Nenek tertidur di beranda.	Odu koodop id pansaran.
23.	Maisara menjadi guru.	Maisara dumadi mongingia.
24.	Hamid belajar Bahasa Kadazandusun.	Hamid balajar boros Kadazandusun.
25.	Semua murid telah mendapat hadiah.	Oinsanai susumikul [telah] nakaanu tutungkap.
26.	Anak itu sangat baik.	Tanak dilo tomod osonong.
27.	Buah pisang itu belum masak lagi.	Tua' pundi dilo' amu ansak kawagu.
28.	Matahari itu sudah tinggi.	Tadau dilo sudah akawas
29.	Langit kemerah-merahan pada waktu senja.	Tawan aragang-ragang ontok timpu minsosodop.
30.	Saya bersekolah di UMS.	Oku poinsikul id UMS.
31.	Budak-budak yang bertempiaran lari itu seperti ditabur air panas.	Tangaanak i minurudsau tangkus dilo' miagal tiasan waig alasu.
32.	Dari Sabah.	Mantad Sabah.
33.	Untuk orang miskin.	Montok tulun mosikin.
34.	Di dalam hutan pagi tadi.	Id aralom gouton kosuabon ditinu.

Due to the Kadazandusun language pattern in VSO, many of the sentence structures are the inverse structure of Malay sentences, where the predicate phrase is in front of the subject phrase. However, there are certain phrases where the structure may be formed using the SVO order. All the translations structure with (Malay SVO sequence) in Table 3 are incorrect based on the sentence order structure, except for the Subject-Verb (SV) structure as shown in examples 8 and 9 respectively (BM: Semua pelajar telah diberi uniform, DMT: Oinsanai susumikul nonuan uniform, and BM: Semua pemimpin masyarakat dijemput hadir, DMT: Oinsanai lalansanon ginumuan alapon rumikot).

Although part of language grammar is being discussed in this paper, it should be noted that the explanations are focused on the capability and output from the preliminary text translation using the algorithm in the DMT system and ways to overcome future problems in the translator. Based on the results in Table 3, two areas will be explored and examined: the word order structure of the translation and the quality of the translation based on the MT perspective using the BLEU score.

3.1. Word Translation Arrangement Rules

First, the investigation of the DMT system from Malay (BM) to Kadazandusun (KD) is looking into the basic grammar concept which is VSO and its extended usages such as VS, VS_{Ap} (Ap is Adverb of place), VSVO, VSOO, VSOV and two types of word order sequence which are similar to the Malay language grammar structure, SVO and SV. To simplify the notation, part of the speech in the sentence structure for each example is omitted.

3.1.1. VSO Order Structure

BM source sentence : Saya menjemur padi.
 KD correct translation : P oposidang oku parai.
 DMT : *Oku poposidang parai.*

BM source sentence : Menteri berkata.
 KD correct translation : Minoboros i montiri.
 DMT : *Montiri minoboros.*

In this structure, the DMT is following the literal translation as provided in the example and also a direct translation of the rule-based word-to-word sequence. In the second sample ('Menteri berkata'), DMT has no '*i*' in the translation. The word '*i*' is known as a ligature or conjunction in certain sentences. Examples of conjunction normally being used in Kadazandusun are '*do*', '*dot*' for connecting a verb and an object, '*di*' for connecting something owned by the owner, and '*ot*' for connecting a count/number phrase and a verb. Future solutions for the DMT algorithm (grammar rule):

- a. Rule 1 for VSO – the verb for Kadazandusun must be placed in the front when the processor encounters the first word (as a verb) in the Malay sentence.
- b. Rule 2 for conjunction – check for a suitable conjunction word.

3.1.2. VS Order Structure

BM source sentence : Nenek tersenyum.
 KD correct translation : Nokongingis i odu.
 DMT : *Odu kongingis.*

DMT output is similar to VSO, where the literal translation is provided and without the use of the conjunction '*i*'. Future solutions for the DMT algorithm (grammar rule) are:

- a. Rule 3 for VS – check for sentences with SV structure in BM and perform VS order.
- b. Rule 2 – update this rule to accommodate VS structure and suitable conjunction.

3.1.3. VSAp Order Structure

BM source sentence : Kami berlari di sekolah.
 KD correct translation : Manangkus yahai hilo id sikul.
 DMT : *Yahai manangkus id sikul.*

Again, the translation by DMT is a literal form. Furthermore, '*hilo*' is not added to the sentence. The word '*hilo*' can be used to point to the location together with '*id*' in Kadazandusun for sentence completeness. Although in normal conversations, '*id*' can be dropped as follows: '*Manangkus yahai hilo sikul*'. Future solution for the DMT algorithm (grammar rule) includes:

- a. Rule 4 for VSAp – extend Rule 3 and check for Ap in the Malay sentence after the verb and add suitable Ap (e.g. '*hilo*') to Kadazandusun translation. Example VS + Ap from sample 2 above is: '*Kongingis i odu hilo id pansaran*'. (Nenek tersenyum di beranda).

3.1.4. VSVO Order Structure

BM source sentence : Saya terlupa memadam api.
 KD correct translation : Nolihuan ku momisok i tapui.
 DMT : *Oku nolihuan momisok tapui.*

Although the semantic meaning of the DMT is not wrong, the output is a literal translation with an incorrect '*oku*' and the conjunction '*i*' was not added to the sentence. Future solutions for the DMT algorithm (grammar rule) for this case are:

- a. Rule 5 for VSVO – extend VSO and VS structure to identify SVVO in Malay sentences.
- b. Rule 2 – add suitable conjunction between VO in the Kadazandusun word order.

3.1.5. VSOO Order Structure

BM source sentence : Saya akan beli ikan untuk Nancy.
 KD correct translation : Bolian ku sada i Nancy.
 DMT : *Oku atantu' boli sada montok Nancy.*

Another drawback of the basic word-to-word sequence translation (without linguistic alignment) in the current DMT system is that it tries to translate every word from the source resulting in an incorrect or unsuitable word selection that may be included in the translation. In a BM sentence, the word 'akan' is translated as 'atantu' in Kadazandusun because the word is simply available in the dictionary. Therefore, in a DMT literal translation, the meaning of the sentence is still understandable, although grammatically wrong. Thus, to upgrade the translation: 1) scan the BM sentence to find the verb (V) to be placed at the beginning of the translation, 2) the conjunction such as 'akan' to the verb brings the meaning of an action to be done, thus, another word translation instead of 'boli' needs to be used ('bolian' in this case, is indicating a future tense 'will buy' or 'akan beli'). In the next part of the sentence, the structure in BM '... ikan untuk Nancy' is Object-Object and KD follows this structure as '... sada i Nancy'. However, the DMT translates the sentence as 'montok Nancy'. Future solutions for the DMT algorithm (grammar rule) are:

- a. Rule 6 – extend Rule 1 (VSO) and Rule 3 (VS) to accommodate the VSOO order.
- b. Rule 2 – extend to check ligature or conjunction in VSOO order.

3.1.6. VSOV Order Structure

BM source sentence	: Kami suruh dia lari. *[to move from the place]
KD correct translation	: Sinuhu dahai isio minogidu.
DMT	: Dahai sinuhu disio tangkus.

In this structure, DMT is able to provide a literal translation. Notably, the meaning of the translation is out of context if the BM sentence is about moving from a certain place, and not 'to run'. In the BM sentence, the subject ('kami' – we) is asking an object/someone ('dia' – he/him) to move, but the translation is vice versa, which is 'he/him' asking 'we/us' to run. In linguistic and grammar contexts, this is a critical translation error. Additionally, if the meaning was 'to run', the literal translation by the DMT is still wrong. Future solution for DMT algorithm (grammar rule):

- a. Rule 6 – extend Rule 1 (VSO) and Rule 3 (VS) to accommodate VSOO order.
- b. Rule 2 – extend to check ligature or conjunction in VSOO order.

3.1.7. SVO Order Structure

BM source sentence	: Semua pelajar telah diberi uniform.
KD correct translation	: Oinsanan tangaanak sikul nonuan do uniform.
DMT	: Oinsanan susumikul telah nonuan uniform.

The structure of the translation follows the SVO order, however, the selection of word pairs may not be correct, because 'pelajar' is translated as 'susumikul' when it should be 'tangaanak sikul'. Moreover, the word 'telah' was not removed from the translation. Future solution for DMT algorithm (grammar rule):

- a. Rule 7 – check for SVO in BM sentence and match with SVO in KD.
- b. All rules – extend the rules to check if certain KD words are a combination of two words in BM or vice versa (e.g., pelajar = 'tanganaak sikul').

3.1.8. SVO Order Structure

BM source sentence	: Semua pemimpin masyarakat dijemput hadir.
KD correct translation	: Oinsanan puru molohingon alapon do rumikot.
DMT	: Oinsanan puru ginumuan alapon rumikot.

Again, the structure of the DMT follows the SV order, however, the selection of words may differ from the correct translation due to word pair availability and algorithm selection in the current system. As an example, 'pemimpin masyarakat' is translated as 'puru ginumuan' while the text in the Kadazandusun news normally uses 'lalansanon mogiigiyon'. This is one such example where a translator from the news company may have used terms in the published news which are different from other human translators.

4. Translation Quality Evaluation using BLEU Score

While there are various MT assessment criteria (summarization, complexity, POS tag, frequency itemset, and association relational item), this initial research focuses on the BLEU score to present an overview of current work. Based on translation quality (without semantic meaning and detailed linguistic analysis), the BLEU score is calculated (according to Equation 1) for each translation given the reference

(correct translation). Table 4 presents the sample evaluation for each sentence. According to the sentence's BLEU score, the average is 0.6735 and about 85% of the sentences have a score above 0.5. With an average of 0.68 BLEU score, the DMT performance (limited to the given sample sentences) has achieved the level of translation that the average performance described in [22]. However, the system still needs to rewrite the sentence structure to match the grammar and semantic meaning. Accordingly, to increase the translation performance, basic sentence structure orders (as identified in the previous section) need to be applied.

Table 4. BLEU score as of the translation system against the human translation reference

No.	Translation using DMT	Expert Human Translation	BLEU Score
1	Montiri minoboros.	Minoboros i montiri.	0.7652
2	Oku poposidang parai.	Poposidang oku parai.	0.9070
3	Odu kongingis.	Nokongingis i odu.	0.5640
4	Yahai manangkus id sikul.	Manangkus yahai hilo id sikul.	0.7372
5	Oku nolihuan momisok tapui.	Nolihuan ku momisok i tapui.	0.8064
6	Oku atantu' boli sada montok Nancy.	Bolian ku sada i Nancy.	0.4361
7	Dahai sinuhu disio tangkus.	Sinuhu dahai isio minogidu.	0.5980
8	Jimmy Palikat tanak kampung.	Tanak kampung i Jimmy Palikat.	0.8243
9	Leha mihad.	Mihad i leha.	0.5533
10	Lihan tomod agahui.	Okugui tomod i lihan.	0.5190
11	Sikul oku id Ranau.	Sekolah saya di Ranau.	0.3521
12	Hilda mongingia'.	Mongingia' i hilda.	0.6644
13	Tapa oku mongingia' tugama.	Mongingia' tugama i tapa ku.	0.6672
14	Tolu tulun mongingia' sikul.	tolu tulun mongingia' sikul.	0.9062
15	Puun piasau.	Puun piasau.	1.0000
16	Tama tapa.	Tina om tapa.	0.4804
17	Kusai tondu.	Kusai om tondu.	0.7192
18	Alice mambasa buuk.	Mambasa I Alice do buuk.	0.6641
19	Halim kongingis.	Kongingis i halim.	0.6943
20	Odu koodop id pansaran.	Kodop i odu hilo id pansaran.	0.6584
21	Maisara dumadi mongingia'.	Dumadi i Maisara do mongingia'.	0.7803
22	Tanak dilo tomod osonong.	Osonong tomod ilo tanak.	0.7450
23	Tua' punti dilo' amu ansak kawagu.	amo po noonsok ilo tua' punti.	0.3891
24	Tadau dilo' sudah akawas	Akawas no ilo tadau.	0.4427
25	Tawan arang-rang ontok timpu minsosodop.	Arang-rang o tawan do minsosodop.	0.7123
26	Oku pomsikul id UMS.	Hilo id UMS oku pomsikul.	0.6957
27	Mantad Kedah.	Mantad Kedah.	1.0000
28	Montok tulun mosikin.	Montok tulun mosikin.	1.0000
29	Id aralom gouton kosuabon ditinu.	Id suang do gouton kosuabon konihab.	0.5513
Average			0.6839

The translation is then tested for a longer text from newspapers with a human translation pair, in this case between Malay (Bernama News) and Kadazandusun (published human translation at New Sabah Times). The sample of the Bahasa Melayu news¹ is shown in Figure 3 and the Kadazandusun translation text adapted from News Sabah Times² is shown in Figure 4. Furthermore, sample translation output from Bahasa Melayu to Kadazandusun using the DMT is presented in Figure 5. The purpose of this translation is to examine the performance of the translation using a relatively long text and the possibility of complex sentences. Evaluating the translation of DMT as compared to the published news, the BLEU score is at 0.8182 (as one whole sentence) and 0.7273 (an average when the sentence is divided into different paragraphs). Again, considering only the given sample, a high BLEU score indicates that the quality of the translation is comparable to human translation. However, it is important to note that BLEU does not evaluate if the translation delivers a similar meaning as the source. Thus, if the semantic meaning is the concern, then other MT evaluation metrics with advanced linguistic analysis should be used.

¹ <https://mediapermata.com.bn/bukit-giling-diwartakan-sebagai-perkampungan/>

² The New Sabah Times news archive is no longer available after closing down since 31 December 2020.

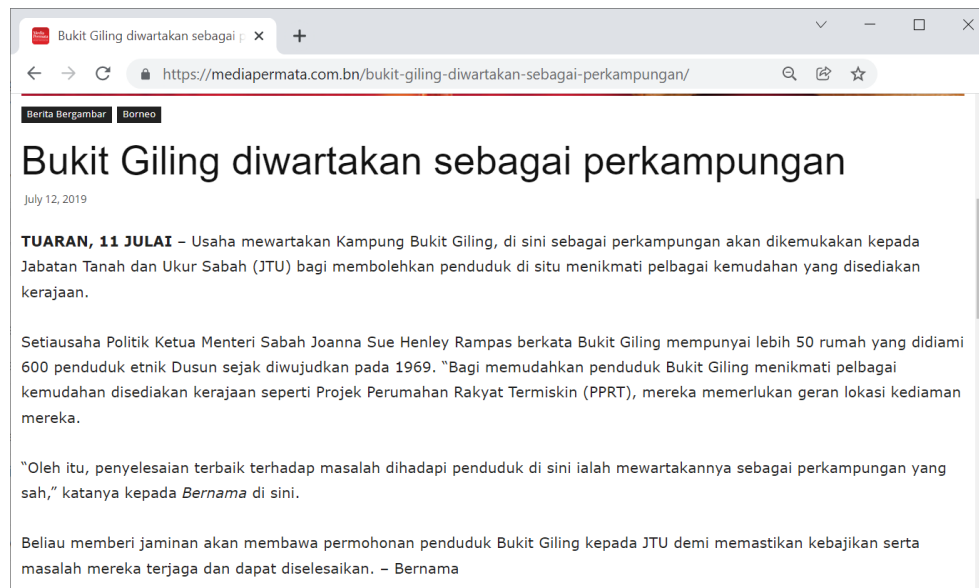


Figure 3. BERNAMA news archive for translation sample

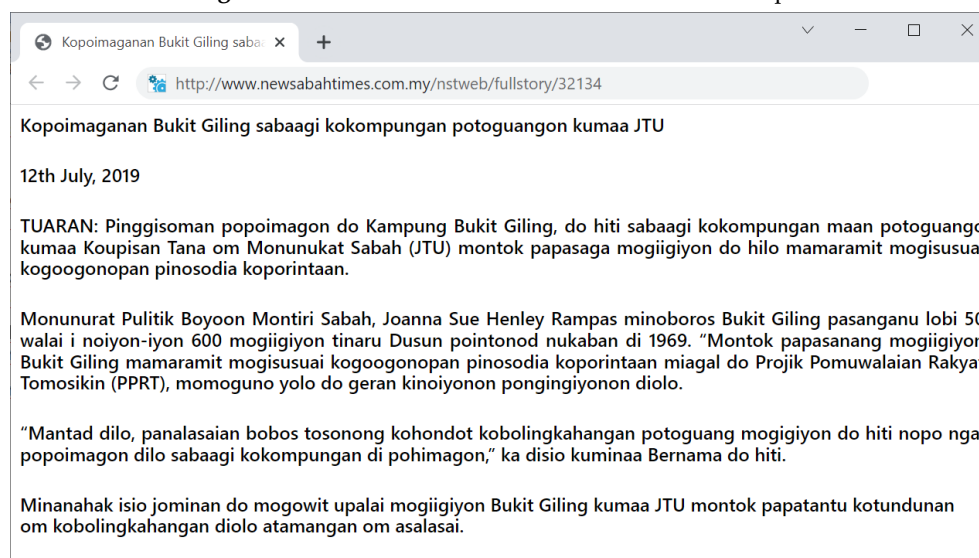


Figure 4. Sample translation of the text in Figure 3 from New Sabah Times news archive

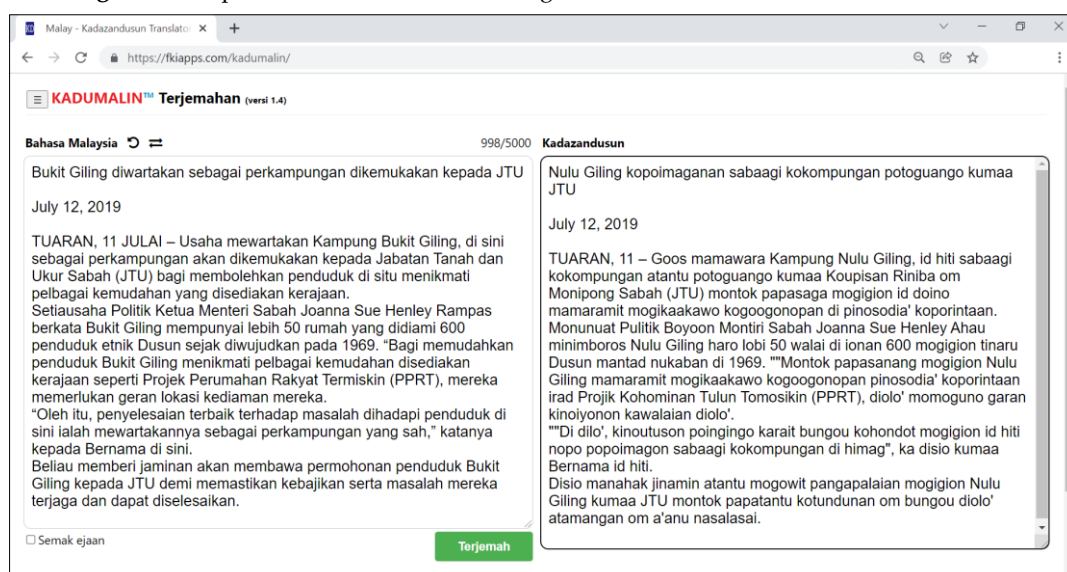


Figure 5. DMT system output to Kadazandusun

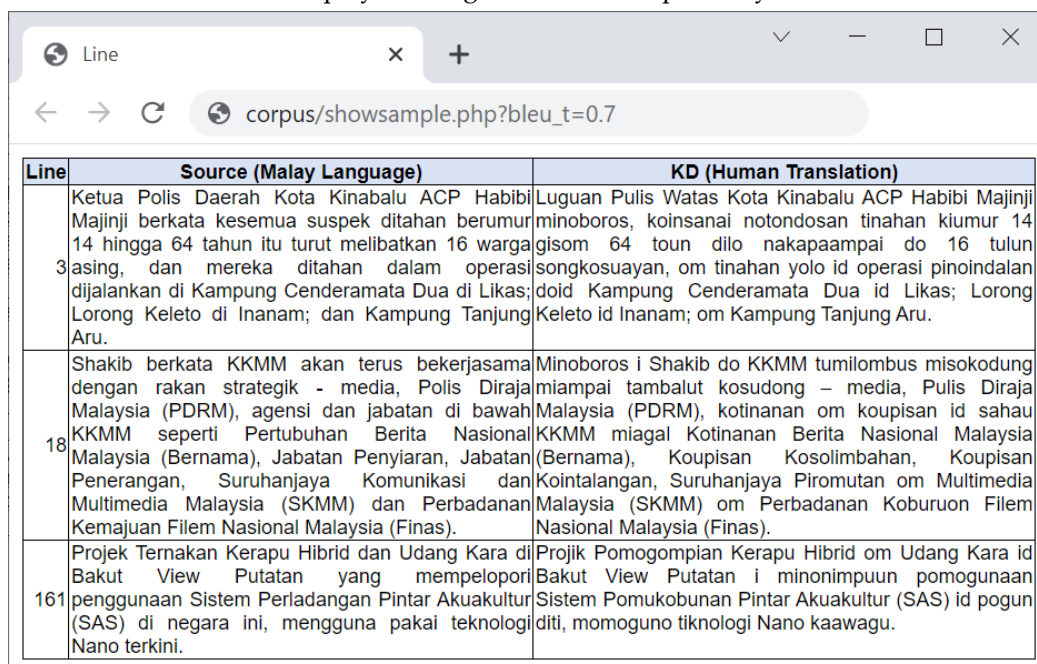
Finally, a longer text corpus from the collection (source indicated in Table 2) was prepared to test the capability of the translation system. A set of reference text and candidate text files were constructed to

evaluate the development of a translation system for Malay-Kadazandusun translation. Both files contain 400 lines of text from Malay (Bernama News) and its Kadazandusun translation (human translation at New Sabah Times). News content in the text ranges from drug-related crime, water tariff, development, health, illegal immigrants, entertainment, events, and politics. The purpose of using a longer text is to investigate the computational load of the language translation system and to explore new words that are not in the database. Based on the prepared text, the summary of information and the results of the translation using the BLEU score are shown in Table 5.

Table 5. Summary of the text collection

Information	Details
Number of lines	400
BM number of words	10204
KD number of words	10813
BM vocabulary size (including person names, places, etc.)	2378
KD vocabulary size (including person names, places, etc.)	2295
KD DMT vocabulary size	2385
BM maximum sentence length	55
KD maximum sentence length	64
KD DMT maximum sentence length	55
KD DMT running time	Approx. 32 seconds
BLEU Score (Average)	0.5221
Sentence translation with BLEU score > 0.5	237
Sentence translation with BLEU score < 0.3	7

Based on Table 5, the number of words and sentence length in KD is always more than its BM pair. This is because some KD translations will add (for example) particles such as *i*, *o*, *do*, *dot*, *no*, *po*, and *nopo nga*. As an example, the word ‘*do*’ in ‘*Pinggisoman popoimagon do Kampung Bukit Giling*’ is a particle added to complete the sentence translation from the Malay sentence ‘*Usahaewartakan Kampung Bukit Giling*’. In contrast, there are some instances where the vocabulary size for BM is more than KD and some of the BM words can be replaced by only a single word in KD. The use of ‘*telah menerima*’ in ‘*Kementerian Pembangunan Luar Bandar Sabah telah menerima laporan*’ is replaced with ‘*nakaramit*’ as shown in the translation ‘*Komontirian Kopotundaan Labus Kakadayan Sabah nakaramit ruputan*’. The DMT maximum sentence length is similar to BM because it translates using a word-to-word method. For reference, Figure 6 lists the samples with high BLEU scores and Figure 7 is the samples with low BLEU scores. The scores are displayed in Figures 8 and 9 respectively.



Line	Source (Malay Language)	KD (Human Translation)
3	Ketua Polis Daerah Kota Kinabalu ACP Habibi Majinji berkata kesemua suspek ditahan berumur 14 hingga 64 tahun itu turut melibatkan 16 warga asing, dan mereka ditahan dalam operasi dijalankan di Kampung Cenderamata Dua di Likas; Lorong Keleto di Inanam; dan Kampung Tanjung Aru.	Luguan Pulis Watas Kota Kinabalu ACP Habibi Majinji minoboros, koinsanai notondosan tinahan kiumur 14 gisom 64 toun dilo nakapaampai do 16 tulun songkosuayan, om tinahan yolo id operasi pinoindalan doid Kampung Cenderamata Dua id Likas; Lorong Keleto id Inanam; om Kampung Tanjung Aru.
18	Shakib berkata KKMM akan terus bekerjasama dengan rakan strategik - media, Polis Diraja Malaysia (PDRM), agensi dan jabatan di bawah KKMM seperti Pertubuhan Berita Nasional Malaysia (Bernama), Jabatan Penyiaran, Jabatan Penerangan, Suruhanjaya Komunikasi dan Multimedia Malaysia (SKMM) dan Perbadanan Kemajuan Filem Nasional Malaysia (Finas).	Minoboros i Shakib do KKMM tumilombus misokodung miampai tambalut kosudong - media, Pulis Diraja Malaysia (PDRM), kotinanan om koupisan id sahuu KKMM miagal Kotinanan Berita Nasional Malaysia (Bernama), Koupisan Kosolimbahan, Koupisan Kointalangan, Suruhanjaya Piromutan om Multimedia Malaysia (SKMM) om Perbadanan Koburuon Filem Nasional Malaysia (Finas).
161	Projek Ternakan Kerapu Hibrid dan Udang Kara di Bakut View Putatan yang mempelopori penggunaan Sistem Perladangan Pintar Akuakultur (SAS) di negara ini, mengguna pakai teknologi Nano terkini.	Projik Pomogomopian Kerapu Hibrid om Udang Kara id Bakut View Putatan i minonimpuun pomogunaan Sistem Pomukobunan Pintar Akuakultur (SAS) id pogun diti, momoguno tiknologi Nano kaawagu.

Figure 6. Sample corpus translation from BM to KD with a high DMT BLEU score

Line	Source (Malay Language)	KD (Human Translation)
81	Jabatan Kesihatan Sabah ambil tindakan segera atasi masalah kebocoran bangunan HQE	Pogonuan laang tosikap monoibau lawasan linimput HQE
268	Sabah tubuh jawatankuasa khas kabinet bangunkan pelancongan luar bandar	Turidongon komitikuasa pinotomod kabinet
292	SESB Jamin Tiada Pengurangan Pekerja	SESB minanahak jominan do aiso' kopongingkurian pakakaraja'
373	DUN Sabah pertama kali bersidang pada 11 Jun selepas Pilihan Raya Umum (PRU) Ke-14, 9 Mei lepas.	DUN Sabah kumoinsan nogi' do pinapaharo pitimbungakan di ko 11tw Ko'onom katalib Pinomilian Tagayo Pogun (PRU) ko-14, ko 9tw Kolimo katatalib.

Figure 7. Sample corpus translation from BM to KD with a low DMT BLEU score

Line	KD DMT Output	BLEU Score
3	Boyoon Pulis Watas Kota Kinabalu ACP Habibi Majinji minoboros kesemua suspek ditahan kiumur 14 hingga 64 toun dilo' turut kapaampai 16 warga asing, om diolo' ditahan aralom operasi dijalankan id Kampung Cenderamata Dua id Likas; Lorong Keleto id Inanam; om Kampung Tanjung Aru.	0.731423
18	Shakib minoboros KKMM atantu' terus bekerjasama miampai rakan strategik - media, Pulis Diraja Malaysia (PDRM), agensi om koupisan id siriba KKMM miagal Pertubuhan Habar Nasional Malaysia (Bernama), Koupisan Penyiaran, Koupisan Penerangan, Suruhanjaya Komunikasi om Multimedia Malaysia (SKMM) om Kotinanan Kinoburuan Filim Nasional Malaysia (Finas).	0.738106
161	Projik Ternakan Kerapu Hibrid om Gipan Kara id Bakut View Putatan i memelopori pomogunaan Sistom Perladangan Pintar Akuakultur (SAS) id pomogunan diti, mengguna pakai teknologi Nano terkini.	0.72638

Figure 8. Sample source (BM), human translation (KD), and KD DMT output with high BLEU scores

In Figure 8, the BLEU score is high despite some of the words not being successfully translated (no word translation pairs in the database). This is because, first, the DMT output has most of the words from human translation regardless of its position. Therefore, to increase the score, words that have no translation pairs in the dictionary need to be added or updated. Secondly, sentences in the samples are in the structure of SVO and SV, except in some parts where there is a structure of VSO. For example, 'Shakib berkata' is translated as 'Minoboros i Shakib', but the DMT output is 'Shakib minoboros'. In another sample, the KD DMT uses different words compared to human selection such as 'Boyoon' instead of 'Luguan' for the word 'Ketua', 'Gipan' over 'Udang' and 'Sistom' for 'Sistem'. Sistom, luguan and gipan are available in the dictionary of Daftar Kata Bahasa Kadazandusun - Bahasa Malaysia [18].

Line	KD DMT Output	BLEU Score
81	Koupisan Kolidasan Sabah anu tindakan segera atasi kobolingkaangan kebocoran bogunan HQE	0.172346
268	Sabah tubuh jawatankuasa poimbida kabinet bangunkan pelancongan labus kakadayan	0.216374
292	SESB Jamin Aiso Pengurangan Kukumaraaja	0.232567
373	DUN Sabah koiso' kali bersidang ontok 11 Jun selepas Pilihan Raya Umum (PRU) Ke-14, 9 Mei lepas.	0.245728

Figure 9. Sample source (BM), human translation (KD), and KD DMT output with low BLEU scores

Low BLEU scores as shown in Figure 9 are news header titles. In news header translation, human translators prefer to condense news headers by removing certain words or rephrasing the headline, resulting in a translation that is different but still provides a similar meaning. For example, the sample in

line 81 of the human translation where 'Jabatan Kesihatan Sabah' is omitted. In sample line 268, the word 'Sabah' was also dropped in the KD translation, therefore not completely translating the whole sentence from the BM headline. In the BM headline, the reason for forming the special committee was given in the sentence, however, this was omitted in the KD translation. The author may have decided to shorten the translation of the headline so readers can read further in the news article. This kind of translation that shortens a sentence is called text paraphrasing or summarizing and is a problem and an important domain in advanced text mining for machine translations with more linguistic power.

In terms of computational load for the system, large text processing requires a longer time to process, in this case, up to 32 seconds for 10,204 words. Compared to Google Translate, which takes less than a second to process the translation but only accepts 5000 characters. A similar length of 5000 characters was also tested with KD DMT and the processing time was also less than a second. Thus, in future translation system development for public use, 5000 characters should be the maximum number of characters per translation task that can be offered for the service. As for the BLEU score, DMT achieved a 0.5 score which is an average translation score that requires a post-editing in the sentences.

5. Conclusion

The Kadazandusun language is considered an under-resourced language due to its limited usage. However, the language is now being taught in selected schools and universities and it is still being used in a local newspaper. In this paper, we discuss the preliminary experiment on the MT for an under-resourced language within Malay-Kadazandusun using direct machine translation or DMT. The findings in this paper agree with the study presented in [23], where challenges for e-translation tools, specifically for contextual understanding and translation quality are imminent. This problem is far more difficult for a language with limited resources where the language preservation activities, experienced practitioners, and software development practitioners are still finding the best form of collaboration. In terms of implementation, the DMT system has been successfully implemented; however, the experiment shows that the current system requires upgrades and development. The current limitations are: 1) the database content of KD words is limited compared to Malay words, 2) KD standard words and their spelling are yet to be confirmed by language experts because the current system uses a vocabulary that was acquired from a KD newspaper (with existing spelling errors) and certain alignments from available KD dictionaries, 3) translation is a rule-based, where there are minimal grammar checks to align the translation according to basic Kadazandusun word arrangement (VSO), 4) checks for word ligatures or conjunctions such as particle *i*, *no*, *po*, *nopo*, and *nopo nga* is implemented in basic rules, and 5) the BLEU score is higher but post-editing is still required due to problems 3 and 4. As mentioned in the Word Translation Arrangement section, there are seven rules suggested to be implemented in future developments of the DMT system specifically to correct word arrangements in the translation and the semantic meaning or equivalence of the sentences. Finally, we hope that this paper will contribute to the future direction of multidisciplinary research in computing, language, and language translation. Additionally, work on deep learning-based machine translation is still being investigated for a similar application.

Acknowledgment

The study is supported under the Universiti Malaysia Sabah grant (SGI0140). Furthermore, this paper is edited and extended from the original version [24] presented at the International Case Study Conference in 2019.

References

- [1] Yin-Lai Yeong, Tien-Ping Tan and Siti Khaotijah Mohammad, "Using Dictionary and Lemmatizer to Improve Low Resource English-Malay Statistical Machine Translation System", *Procedia Computer Science*, Vol. 81, pp. 243–249, 2016, Online ISSN: 1877-0509, Published by Elsevier B.V., DOI: 10.1016/j.procs.2016.04.056, Available: <https://www.sciencedirect.com/science/article/pii/S1877050916300709>.

- [2] Tong Loong-Cheong, "English-Malay translation system: a laboratory prototype", in *Proceedings of the 11th Conference on Computational Linguistics*, Bonn, Germany, August 1986, pp. 639–642, DOI: 10.3115/991365.991552, Available: <https://dl.acm.org/doi/pdf/10.3115/991365.991552>.
- [3] Tong Loong-Cheong, "The Computer Translation of Interrogatives from English to Malay", *RELC Journal*, Vol. 18, No. 1, pp. 1–18, 1987, Published by SAGE Publications, DOI: 10.1177/003368828701800101, Available: <https://journals.sagepub.com/doi/abs/10.1177/003368828701800101>.
- [4] Kentaro Ogura, Francis Bond and Yoshifumi Ooyama, "A prototype Japanese-to-Malay Translation System", in *Proceedings of the MT Summit VIII*, 13-17 September 1999, Singapore, pp. 444-448, Available: <https://aclanthology.org/1999.mtsummit-1.66/>.
- [5] Pidong Wang, Preslav Nakov and Hwee Tou Ng, "Source Language Adaptation Approaches for Resource-Poor Machine Translation", *Computational Linguistics*, Vol. 42, No. 2, pp. 277–307, 2016, DOI: 10.1162/COLI_a_00248, Available: https://dl.acm.org/doi/10.1162/COLI_a_00248.
- [6] Ahmed Jumaa Alsaket and Mohd Juzaidin Ab Aziz, "Arabic-malay machine translation using rule-based approach", *Journal of Computer Science*, Vol. 10, No. 6, pp. 1062–1068, 2014, DOI: 10.3844/jcssp.2014.1062.1068, Available: <https://thescipub.com/pdf/jcssp.2014.1062.1068.pdf>.
- [7] Hamida Ali Almshrky and Mohd Juzaidin Ab Aziz "Arabic Malay Machine Translation for a Dialogue System", *Journal of Applied Sciences*, Vol. 7, pp. 1371–1377, 2012, DOI: 10.3923/jas.2012.1371.1377, Available: <https://scialert.net/fulltext/?doi=jas.2012.1371.1377>.
- [8] Surafel Melaku Lakew, Aliia Erofeeva and Marcello Federico, "Neural Machine Translation into Language Varieties", in *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT 2018)*, Belgium, Brussels, October 31 - November 1 2018, Published by Association for Computational Linguistics, ISBN 978-1-948087-81-0, DOI: 10.18653/v1/W18-6316, Available: <https://aclanthology.org/W18-6316/>.
- [9] Chong Chai Chua, T. Lim, Lay-Ki Soon, E. Tang and Bali Ranaivo-Malançon, "Analogical-Based Translation Hypothesis Derivation with Structural Semantics for English to Malay Example-Based Machine Translation", *Advanced Science Letters*, Vol. 24, No. 2, pp. 1263–1267, 2018, DOI: 10.1166/asl.2018.10729, Available: <https://www.ingentaconnect.com/contentone/asp/asl/2018/00000024/00000002/art00103>.
- [10] John Oladosu, Adebimpe Esan, Ibrahim Adeyanju, Benjamin Adegoke, Olatayo Olaniyan and Bolaji Omodunbi, "Approaches to Machine Translation: A Review", *FUOYE Journal of Engineering and Technology*, Vol. 1, No. 1, pp. 120–126, 30th September 2016, Published by Federal University of Oye-Ekiti, DOI: 10.46792/fuoyejt.v1i1.26, Available: <http://journal.engineering.fuoye.edu.ng/index.php/engineer/article/view/26>.
- [11] S. Anbukkarasi and S. Varadhaganapathy, "Machine Translation (MT) Techniques for Indian Languages", *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 8, No. 2S4, pp. 86–90, 2019, ISSN: 2277-3878, Published by Blue Eyes Intelligence Engineering & Sciences Publication, DOI: 10.35940/IJRTE.B1015.0782S419, Available: <https://www.ijrte.org/wp-content/uploads/papers/v8i2S4/B10150782S419.pdf>.
- [12] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania, USA, 7-12 July 2002, pp. 311–318, DOI: 10.3115/1073083.1073135, Available: <https://dl.acm.org/doi/10.3115/1073083.1073135>.
- [13] George Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics", in *Proceedings of the second international conference on Human Language Technology Research (HLT'02)*, 24-27 March 2002, San Diego, California, pp. 138–145, Published by Morgan Kaufmann, DOI: 10.3115/1289189.1289273, Available: <https://dl.acm.org/doi/10.5555/1289189.1289273>.
- [14] Michael Denkowski and Alon Lavie, "Meteor universal: language specific translation evaluation for any target language", in *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, June 2014, pp. 376–380, DOI: 10.3115/v1/W14-3348, Available: <https://aclanthology.org/W14-3348/>.
- [15] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, "Perplexity measure of the difficulty of speech recognition tasks", *The Journal of the Acoustical Society of America*, Vol. 62, No. S1, pp. S63, 1977, DOI: 10.1121/1.2016299, Available: <https://asa.scitation.org/doi/10.1121/1.2016299>.
- [16] Francisco Guzmán, Shafiq Joty, Lluís Màrquez and Preslav Nakov, "Machine translation evaluation with neural networks", *Computer Speech & Language*, Vol. 45, pp. 180–200, 2017, DOI: 10.1016/j.csl.2016.12.005, Available: <https://www.sciencedirect.com/science/article/abs/pii/S0885230816301693>.
- [17] Karunesh Kumar Arora and Shyam S. Agrawal, "Pre-Processing of English-Hindi Corpus for Statistical Machine Translation", *Computación y Sistemas*, Vol. 21, No. 4, 2017, pp. 725–737, DOI: 10.13053/CyS-21-4-2697, Available: <https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2697>.
- [18] Kadazandusun Language Foundation, *Daftar Kata Bahasa Kadazandusun - Bahasa Malaysia*. Kadazandusun Language Foundation (KLF) and INDEP Education Foundation (IEF), ISBN: 978-983-9325-64-5, 2015.
- [19] Ree John Daulip, *Kamus Malay-Dusun-English*. 2nd Ed., Kota Kinabalu, Sabah: Sabah Komik, ISBN: 978-967-11632-2-1, 2015.

- [20] Kementerian Pendidikan Malaysia, *Puralan boros Kadazandusun id sikul*. Putrajaya: Bahagian Pembangunan Kurikulum, ISBN: 9789675094095, 2008.
- [21] Minah Sintian, "Struktur binaan ayat Bahasa Kadazandusun dan Bahasa Melayu: satu pengenalan", in *Proceedings of the Seminar Antarabangsa Susastera, Bahasa dan Budaya Nusantara*, 2019, Online ISBN: 978-967-0922-79-9, pp. 215-228, Available: <http://dspace.unimap.edu.my/xmlui/handle/123456789/69123>.
- [22] Seamus Lyons, "Quality of Thai to English Machine Translation", *Knowledge Management and Acquisition for Intelligent Systems, Lecture Notes in Computer Science Book Series*, vol. 9806, pp. 261–270, 2016, DOI: 10.1007/978-3-319-42706-5_20, Available: https://link.springer.com/chapter/10.1007/978-3-319-42706-5_20.
- [23] Dianne Excell, "Some Problems in Using Computer-Aided Translation Tools to Facilitate Second Language Fluency in Education", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 22-31, Vol. 3, No. 2, 1st April 2019, Published by International Association for Educators and Researchers (IAER), DOI:10.33166/AETiC.2019.02.003, Available: <http://aetic.theiaer.org/archive/v3/v3n2/p3.html>.
- [24] Mohd Shamrie Sainin, Mohammad Zulfarhan Humin, Asni Tahir and Suraya Alias, "Machine Translation: Case Study for Kadazandusun Text Translation", in *Proceedings of the 4th International Case Study Conference*, Sabah, Malaysia, 24-26 November 2019, E-ISBN:978-967-11030-8-1, pp. 250-257, Published by, Universiti Utara Malaysia, Available: http://www.imbre.uum.edu.my/media/attachments/2021/02/07/proc_icsc2019-compressed.pdf.



© 2023 by the author(s). Published by Annals of Emerging Technologies in Computing (AETiC), under the terms and conditions of the Creative Commons Attribution (CC BY) license which can be accessed at <http://creativecommons.org/licenses/by/4.0>.