Statistical Machine Translation Dayak Language – Indonesia Language

Muhammad Fiqri Khaikal [1), Arie Ardiyanti Suryani [2)

Informatika, Fakultas Informatika, Universitas Telkom
Jl.Telekomunikasi Terusan Buah Batu Sukapura Kec. Dayeuhkolot, Kota Bandung, 40257
e-mail : fiqrikhaikal@student.telkomuniversity.ac.id [1); ardiyanti@telkomuniversity.ac.id [2)

ABSTRACT

This Paper aims to discuss how to create the local language machine translation of Indonesia Language where the reason of local language selection was carried out as considering the using of machine translator for local language are still infrequently found mainly for Dayak Language machine translator. Machine Translation on this research had used statistical approach where the resource data that was taken originated from articles on dayaknews.com pages with total parallel corpus was approximately 1000 Dayak Language – Indonesia Language furthermore as this research contains the corpus with total 1000 sentences accordingly divided into three sections in order to comprehend the certain analysis from a pattern that was created. The monolingual corpus was collected approximately 1000 sentences of Indonesia Language. The testing was carried out using Bilingual Evaluation Understudy (BLEU) tool and had result the highest accuracy value amounting to 49.15% which increase from some the others machine translator amounting to approximately 3%.

Key words - Machine Translator, Statistical, Parallel Corpus, Monolingual Corpus, BLEU

1. INTRODUCTION

A. Background of Study

Language as a tool in communication used as an interaction medium between individual, Indonesia has a large diversity in local language, the Dayak local language is one that predicted extinct in 20 - 30 years ahead as parent didn't teach the local language since childhood, however parent have been teaching Indonesia Language as said by Hery Budhiono from Central Kalimantan Language Institution. Various efforts have been attempted to save the local languages which tend to lead to the extinction process(Ansori, 2019). For the example local language become one of the local content subject in the elementary school in certain area, conducting research and seminars from time to time, the University Level have been opening a study program or regional language and literature department, however all those efforts are not appropriate solution to survive local language from extinct. (Darwis, 2011).

The machine translation which capable to translate from Indonesia Language to Local Language is created to solve the problems. The creating of machine translation has the aim to facilitate difference in communication between various individual, generally in local language and specifically Dayak Language to avoid from extinct and could be sustained. This research has been using statistical approach, Machine Translation as paradigm for Machine Translation wherein the translation was translated in statistic model with parameters lowered from the corpus parallel analysis (Hadi, 2014).

Nowadays machine translation for Indonesian Language - Dayak Language is rarely found. Some research have been discussing regarding statistical machine translation, Machine Translation Indonesia Language – Bugis Wajo Language was written by Mulyana dkk that has accuracy level 59.19%(Mulyana et al., 2018), Indonesia Language – Dayak Kanayat

has accuracy average 50.14% (Sujaini, 2017). Local Language Statistical Machine Translation are rarely found, therefore this thesis contribute the statistical machine translation for Indonesian Language – Dayak Language to enrich statistical Machine Translation using local language.

B. Topic and Discussion

Some issues discuss on this paper were how to build and develop Dayak Language - Indonesia Language statistical machine translation and figuring out the accuracy value created from Dayak Language - Indonesia Language Statistical Machine Translator.

The scope of problem limitation on this paper for parallel corpus were collected from a website called Dayaknews.com with total collected parallel corpus limited to approximately 1000 lines of sentences and translation just for one way only which was Indonesia Language as target language and Dayak Language as resource language.

C. Goal

This research aims to generate a statistical machine translation using Indonesia Language - Dayak Language basis completed with good translation accuracy and facilitate translation easily for learner Dayak Language and to increase the resources of the local language translation machine.

2. LITERATURE REVIEW

A. Statistical Machine Translation

Statistical Machine Translation is a paradigm from automatic translation which is done with a statistics-based approach(Hadi, 2014). Inside statistical machine translation, translation features are taken from parallel corpus and monolingual corpus entered during training. Architecture within statistical machine translation as shown in the following figure no. 1.
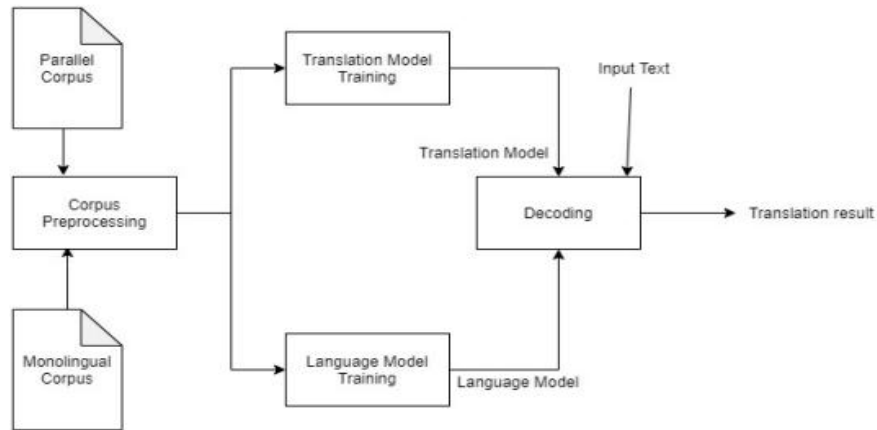
Figure 1. Statistic Machine Translation Architecture.

Figure 1 was the architecture components from statistical machine translation consists of parallel corpus, monolingual corpus, preprocessing corpus, language models, translation models and decoder.

Parallel corpus and monolingual corpus as resource languages that have been made then processed in the preprocessing corpus step. Language model aims to generate the probability sentence that will appear and translation model function to divide original language sentence become phrase line, translate each phrase to its destination and recording.

Decoder functions in finding text to the most probability destination language with considering some various components such as language model and translation model.(Asparilla et al., 2018).

B. Language Model (LM)

Language Model functions to counting probability from the sentences that possibly appear. Before counting the sentences probabilities, firstly need to count the word probabilities as word order comes first than sentence order with chain rule or commonly called n-gram model formula as following (Manual & Guide, 2012):

$$P(K) = P_{(W1,W2....Wn)} \dots\dots\dots\dots\dots\dots\dots\dots\dots(1)$$

Where K is sentence notation and W   is word notation. N-gram language model consist of three languages: unigram, bigram and trigram. Unigram is the emerging of word that is not affected by other words. Bigram is the emerging of word that is affected by another word. Trigram is the emerging of word that is affected by the previous word.

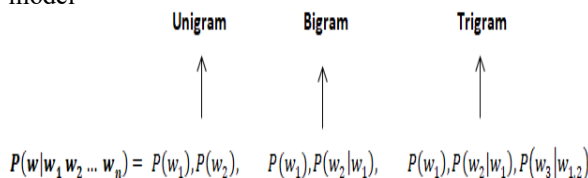Following are the three of n-gram language model

Unigram   Bigram   Trigram

$$P(w|w_1 w_2 \dots w_n) = P(w_1),P(w_2), \quad P(w_1),P(w_2|w_1), \quad P(w_1),P(w_2|w_1),P(w_3|w_{1,2})$$

Figure 2. N-gram language model

C. Translation Model(TM)

Translation model has function to determine and finding the accuracy of the word. Translation model pair the input text in source language and output text in target language.

Translation model has two types of methods in statistic approach: phrase-based approach (phrase-based translation) and word-based approach (Word-based translation).   Phrase-based approached was applied in this research as considering that this approach more capable to resolve the ambiguity problem compared to word-based approach as well as compared to differential evolution approach, phrase-based approach is more effective (Dugonik et al., 2015).

Some major steps in phrase-based translation model:

1. Creating the source sentence into a phrase table
2. Translate every phrase into target language.
3. Recording step.

D. Corpus

According to Baker (2010:93) corpus is collection of some texts both written and oral that are stored in computer.   Baker defines that corpus available in electronic media only. Unlike according to Setiawan (2017) corpus is a writing that written by someone in the form of hard copy of soft copy such as book, magazine, dictionary and newspaper, whereas soft copy such as application, website, online dictionary etc.

E. Bilingual Evaluation Understudy (BLEU)

Bilingual evaluation understudy (BLEU) is an algorithm that functions in evaluating the quality of translation translated by machine from natural language to another language. BLEU measures a modified n-gram precision score between automatic translation results with reference translation using constant called brevity penalty. (Wentzel, 1922).

BLEU score was obtained from multiplication between brevity penalty with geometric average from

modified precision score. Higher the BLEU score, then more accurate with the reference. It is very important to know that more reference of translation per sentence, the score will higher accordingly. In order to obtain the high BLEU score, the translation sentence length must be close to the reference of sentence length and the sentence must have the same word and in sequence with reference sentence. The following is BLEU formula:

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \quad\dots\dots\dots\dots\dots\dots\dots(2)$$

$$P_n = \frac{\sum_{C \in corpus} \sum_{n-gram\ \in c} count_{clip}(n-gram)}{\sum_{C \in corpus} \sum_{n-gram\ \in c} count\ (n-gram)}\dots(3)$$

$$BLEU = BP.\exp\left(\sum_{n-1}^{N} w_n log p_n\right) \quad\dots\dots\dots\dots(4)$$

BP = brevity penalty, is the candidate penalty during the translation of sentence. (c) Longer than the reference (r)

C   = Total word resulted from automatic translation
r   = the total of reference word.
r   = modified precision score
W   = (standard N value for BLEU is 4)
Pn   = the total of n-gram resulted from translation suitable with reference divided with total n-gram resulted from translation.

3. RESEARCH METHODOLOGY

Herein was research methodology for statistical machine translation Dayak Language – Indonesia Language that will be applied in this thesis wherein some process, as follow:
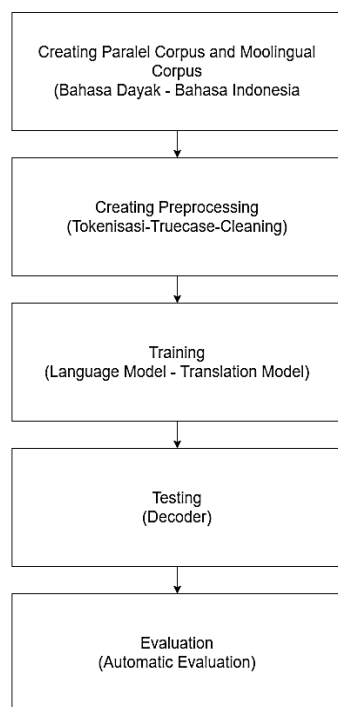

Figure 3. Design system of Statistical Machine Translation Indonesia Language - Dayak Language

A. Monolingual Corpus and Paralel Corpus

On this thesis, the main data required in creating the statistical machine translation was parallel corpus and monolingual corpus. The source of the corpus used was taken from the website called Dayaknews.com. Parallel corpus consist of approximately 1000 lines of sentences of Indonesia Language and Dayak Language, whereas Monolingual corpus consist of approximately 1000 lines of sentences Indonesia Language.

B. Corpus Preprocessing

The preprocessing corpus step aims to prepare a corpus. Parallel corpus and monolingual corpus will processed using tool called MOSES, the tool functions to train parallel corpus and monolingual corpus become translation and model of language(Mandira et al., 2016) and had some steps such as tokenization, true case and cleaning.

1. Tokenisasi is a spacing process between word and spacing process between word and punctuation. Example. Tokenization of comma in Dayak Language:

Input sentence = putiki jemoran, dah merian ni.
Tokenization result =
- putiki jemoran
- dah merian ni.

The tokenization command can be viewed in the fig. 4

Figure 4. Tokenization Command from Machine 1

2. Truecase is a process that aims to convert the prefix every sentence according to the most possibilities place. Truecase is divided into 2 sections: uppercase and lowercase. Herein is the example of command to create a command from truecase that can be viewed in fig. 5.

Figure 5. Truecase Command from Machine 1

3. Cleaning is the process of removing white spaces and limiting sentence length. The limitation of sentence length is to be done in order to reduce error prone and accuracy of a translation that aim to maintain the quality of the translation results. The example of cleaning command can be viewed in fig. 6.

Figure 6. Cleaning Command from Machine 1

The fig. 6 is cleaning process of a very long sentence wherein the maximum sentence length is 80 words. If exceeds then sentence will be deleted.

C. Training

On the training step parallel corpus and monolingual corpus will be processed to generate a language model and translation model. Language model aim to obtain the model from target language that is Indonesia Language, whereas the translation model aims to generate model from translation using phrase-based approach. The language model is run using a tool called SRILM (the SRI Language Modeling). This tool is used for n-gram based modeling, whilst translation model is run using tool called GIZA++. Following is the example of training command.

```
fiqrikhaikal@fiqrikhaikal-VirtualBox:~/tafiqri/smt/working1$ nohup nice /home/fi
qrikhaikal/tafiqri/smt/scripts/training/train-model.perl -root-dir train -corpus
 /home/fiqrikhaikal/tafiqri/smt/corpus/corpus1/corpus311.clean -f dyk -e idn -al
ignment grow-diag-final-and -reordering msd-bidirectional-fe -lm 0:3:/home/fiqri
khaikal/tafiqri/smt/corpus/corpus1/LM/corpusmono.blm:8 -external-bin-dir /home/f
iqrikhaikal/tafiqri/smt/tools >& training.out &
[1] 2508
```

Figure 7. Training Command from Machine 1

The fig. 7 is training process to build statistic machine translation and generate the document called moses.

D. Testing

In the testing step using decoder in the tool MOSES which aims to translate the source language to the target language. The decoder works to find the text in the target language that has the biggest probability with several comparisons between the language model and the translation model.

Moses Decoder translates an input sentence from source language that is Indonesia Language. Hereinafter the input sentence is processed using moses decoder, and generate an output sentence in the form of target language that is Dayak Language. The following is command from testing run by moses:

```
fiqrikhaikal@fiqrikhaikal-VirtualBox:~/tafiqri/smt/working1$ /home/fiqrikhaikal/
tafiqri/smt/bin/moses -f /home/fiqrikhaikal/tafiqri/smt/working1/binarised-model
/moses.ini
```

Figure 8. Decoding Command from Machine 1

E. Evaluation

In the last step, there is an evaluation result using the testing step result using BLEU software. Following is command from evaluation process using BLEU can be viewed in the fig. 9:

```
fiqrikhaikal@fiqrikhaikal-VirtualBox:~/tafiqri/smt$ /home/fiqrikhaikal/tafiqri/s
mt/scripts/generic/multi-bleu.perl -lc /home/fiqrikhaikal/tafiqri/smt/corpus/cor
pus1/testing/corpus-testing.true.idn < /home/fiqrikhaikal/tafiqri/smt/working1/c
orpus311.translated.idn
BLEU = 43.76, 68.1/49.1/37.5/29.2 (BP=1.000, ratio=1.086, hyp_len=7108, ref_len=
6548)
```

Figure 9. The command to calculate the BLEU score from the machine 1

4. RESULT AND DISCUSSION
A. Testing Result

The following are some of the test results carried out by a statistical machine translation Dayak Language – Indonesia Language can be viewed in the table 1.

Tabel 1. Hasil dari Score BLEU

	Machine Translation Configuration 1	Machine Translation Configuration 2	Machine Translation Configuration 3
Number of Parallel Corpus (Training)	311	504	806
Number of Monolingual Corpus	1164	1164	1164
Number of Corpus Data Testing (Testing)	362	362	362
BLEU Score Result (%)	43,76%	46,15%	49,15%
Percentage of Increase (%)		+2,39%	+3%

Table 1 contains the total of evaluation from machine translation 1 – 3, wherein the Machine 1 had resulted BLEU score 46.15%, Machine 2 had resulted BLEU score 46.15% which increased 2.39% and Machine 3 had resulted BLEU score 49.15% and increased 39%.

Furthermore from all three machine translation, one machine with the highest BLEU score was taken that was machine no.3 then retested using some different monolingual corpus to add the analysis from machine translation that will be created. Following table 2 shown results from testing.

Tabel 2. Hasil dari Score BLEU

	Machine Translation Configuration 3	Machine Translation Configuration 4	Machine Translation Configuration 5
Number of Parallel Corpus (Training)	806	806	806
Number of Monolingual Corpus	1164	770	392

	Machine Translation Configuration 3	Machine Translation Configuration 4	Machine Translation Configuration 5
Number of Corpus Data Testing (Testing)	362	362	362
BLEU Score Result (%)	49,15%	49.01%	48.80%
Percentage of Decrease (%)		-0,14%	-0,21%

Table 2 contains the total of previous evaluation wherein the machine 4 used monolingual corpus less than machine 3 and had resulted the BLEU score 49.01% which decreased 0.14%, machine 5 with monolingual corpus with total 392 had resulted BLEU score 48.80% and decreased 0.21%.

Herein was implementation result from statistical machine translation Dayak Language - Indonesian Language in website-based. The following were some of the displays in the user interface that have been created

Figure 10. Main Page (Input text)

In the figure 10 there was an initial display on the statistical translation machine Dayak Language - Indonesia Language where the display located on the text input page which has several functions such as:
1. Select the configuration of machine translation
2. Enter the word / sentence in Dayak Language.
3. Translate the word / sentence of Dayak Language.

After saw the initial display of the text input page. The following was a display at the testing stage of the text input page which can be seen in Figure 11.

Figure 11. Testing View (*input text*)

Figure 11 shown the display of the test in the previous figure where the user selects configuration 3 with an accuracy of 49.15% and enters a sentence then translated.

Next there was a display of the initial page of file upload which can be seen in Figure 12.

Figure 12. Main Page (upload file)

Figure 12 there was the initial page display from file upload page wherein the page has some functions such as :
1. Select the configuration of machine translation
2. Uploading a file of resource language for Dayak language which will be translated into Indonesia Language.
3. Uploading a reference file of Indonesia Language
4. Comparing both resource file and reference file.

After saw the initial display of file upload page, the following was the display at the testing stage wherein before testing, username and password must be entered so that it can be accessed from the file upload page which can be seen in Figure 13.

Figure 13. Testing view (upload file)

Figure 13. shown the test results on the file upload page wherein the testing selects configuration 3 and uploads a source file "Corpus-testing true dyk" and a reference file "corpus-estng true idn" and resulted a BLEU score of 49.15%.

B. Discussion of Testing Results

The followings were discussion on the testing results that was carried out.

1. Automatic evaluation of the translation results all corpuses on the statistical translation machine of the Dayak language - Indonesia language using training corpus of 3111 lines of sentences and 11644 monolinguals lines resulting in a BLEU score of 43.76%. the translation results of entire corpus with training corpus 504 lines of sentences and monolingual 1164 lines of sentences had resulted BLEU score 46,15%. whilst the translation results of entire corpus with training corpus 806 lines of sentences and monolingual 1164 lines of sentences resulted BLEU score 49.15%. For the corpus using training corpus 806 lines of sentences and monolingual corpus 770 lines of sentences resulted BLEU score 49.01% and for corpus using monolingual corpus 392 sentences lines with training corpus 806 lines of sentences resulted BLEU score 48.80%.

2. The increasing in the BLEU score for machine translation 1 against the translation machine 2 with a difference in corpus training 193 lines of sentences had increased + 2.39% then from translation machine 2 to translation machine 3 with a difference in corpus training 302 lines of sentences had also increased in the BLEU score + 3% which can be seen from differences in several sentences that exist in several cases which were as follows:

Case number-1
Input Sentence :
"Karena kegiatan jituh inggawi huang bulan puasa, maka hidangan lauk jituh akan menjadi ije sajian buka puasa hayak-hayak," kuan iye.
Machine Translation Configuration 1 :
"Karena kegiatan tersebut dilakukan pada bulan puasa, maka lauk hidangan ini akan menjadi salah satu sajian buka puasa bersama," kata dia.
Machine Translation Configuration 2 :

"Karena kegiatan ini dilakukan pada bulan puasa, maka lauk hidangan ini akan menjadi salah satu sajian buka puasa royong," katanya.
Machine Translation Configuration 3 :
"Karena kegiatan ini dilakukan pada bulan puasa, maka lauk hidangan ini akan menjadi salah satu sajian buka puasa bersama," katanya.
Target Sentence
"Karena kegiatan ini dilaksanakan pada bulan puasa, maka hidangan ikan ini akan menjadi salah satu sajian buka puasa bersama," jelasnya.
Case number-2
Input Sentence :
Dino (30) uluh tebela bara Palangka Raya ji wayah tuh melai hong Kuala Kurun Kabupaten Gunung Mas (Gumas), harus berurusan dengan Polres Gumas, karena komentarnya hong media sosial, andau Minggu.
Machine Translation Configuration 1 :
Dino (30) Pemuda dari Palangka Raya yang saat ini melai di Kabupaten Kuala Kurun Gunung Mas (Gumas), harus berurusan dengan Polres Gumas, karena komentarnya di media sosial, hari Minggu.
Machine Translation Configuration 2 :
Dino (30) orang muda dari Palangka Raya yang kini berada di Kabupaten Kuala Kurun Gunung Mas (Gumas), harus berurusan dengan Polres Gumas, karena komentarnya di media sosial, hari Minggu.
Machine Translation Configuration 3 :
Dino (30) pemuda dari Palangka Raya yang kini tinggal di Kabupaten Kuala Kurun Gunung Mas (Gumas), harus berurusan dengan Polres Gumas, karena komentarnya di media sosial, Minggu.
Target Sentence
Dino (30) pemuda asal Palangka Raya yang sekarang tinggal di Kuala Kurun Kabupaten Gunung Mas (Gumas), harus berurusan dengan Polres Gumas, karena komentarnya di media sosial, Minggu.

Case number-3
Input Sentence :
Kapolres Gumas AKBP Rudi Asriman, S.I.K., mengutak, Humas Polres Gumas hayak dengan Satreskrim jadi mengetawa unggahan tersewut langsung mentehau dan melakukan pemeriksaan dengan pelaku.
Machine Translation Configuration 1 :
Kapolres Gumas AKBP Rudi Asriman, S.I.K., mengatakan, Humas Polres Gumas sama dengan Satreskrim sudah mengetawa unggahan tersebut langsung panggilan dan melakukan pemeriksaan dengan pelaku.
Machine Translation Configuration 2 :
Kapolres Gumas AKBP Rudi Asriman, S.I.K., mengatakan, Humas Polres Gumas sama dengan Satreskrim sudah mengetahui unggahan tersebut langsung panggilan dan melakukan pemeriksaan dengan pelaku.
Machine Translation Configuration 3 :
Kapolres Gumas AKBP Rudi Asriman, S.I.K., mengatakan, Humas Polres Gumas bersama dengan Satreskrim sudah mengetahui unggahan tersebut langsung panggilan dan melakukan pemeriksaan dengan pelaku.

Target Sentence
Kapolres Gumas AKBP Rudi Asriman, S.I.K., mengatakan, Humas Polres Gumas bersama dengan Satreskrim setelah mengetahui unggahan tersebut langsung memanggil dan melakukan pemeriksaan kepada pelaku.

Case number-4
Input Sentence :
Namun hong eka kejadian jadi tege sekitar 10 biti ji sedang menunggu ji dia ' kesene awi korban, dan ije biti pelaku TR misek dengan korban tuntang chat.
Machine Translation Configuration 1 :
Namun di tempat kejadian sudah ada sekitar 10 orang yang sedang menunggu yang tidak kesene oleh korban, dan salah satu pelaku TR misek dengan korban sehingga chat.
Machine Translation Configuration 2 :
Namun di tempat kejadian sudah ada sekitar 10 orang yang sedang menunggu yang tidak kesene oleh korban, dan salah satu pelaku TR misek dengan korban sehingga chat.
Machine Translation Configuration 3 :
Namun di tempat kejadian sudah ada sekitar 10 orang yang sedang menunggu yang tidak kesene oleh korban, dan salah-seorang pelaku TR misek dengan korban sehingga chat.
Target Sentence
namun di lokasi kejadian sudah ada sekitar 10 orang yang sedang menunggu yang tidak dikenal oleh korban, dan salah seorang pelaku TR mempertanyakan kepada korban tentang chat.

Case number-5
Input Sentence :
tege kea pire-pire menu khas Dayak.
Machine Translation Configuration 1 :
ada juga pire-pire menu khas Dayak.
Machine Translation Configuration 2 :
ada juga beberapa menu khas Dayak.
Machine Translation Configuration 3 :
ada juga beberapa menu khas Dayak.
Target Sentence
ada pula beberapa menu khas Dayak.

Case number-6
Input Sentence :
rega ji terjangkau hong kantong uluh tebela nempara bara Rp 7 ribu hingga Rp 20 ribu.
Machine Translation Configuration 1 :
rega yang terjangkau di kantong Pemuda nempara dari Rp 7 ribu hingga Rp 20 ribu.
Machine Translation Configuration 2 :
harga yang terjangkau di kantong orang muda mulai dari Rp 7. hingga Rp 20.
Machine Translation Configuration 3 :
harga yang terjangkau di kantong pemuda mulai dari Rp 7 ribu hingga Rp 20 ribu.
Target Sentence
harga yang terjangkau dikantong kawula muda mulai dari Rp 7 ribu hingga Rp 20 ribu.

Case number-7
Input Sentence :
penginan unggulan sukses hantarkan Kabupaten Kapuas juara ije se-Kalteng.
Machine Translation Configuration 1 :
makanan unggulan sukses hantarkan Kabupaten Kapuas juara satu se-Kalteng.
Machine Translation Configuration 2 :
kuliner unggulan sukses hantarkan Kabupaten Kapuas juara satu se-Kalteng.
Machine Translation Configuration 3 :
kuliner unggulan sukses hantarkan Kabupaten Kapuas juara satu se-Kalteng.
Target Sentence
kuliner unggulan sukses hantarkan Kabupaten Kapuas juara satu se-Kalteng.

Case number-8
Input Sentence :
"Jadi, itah beharap dengan tege Bumdesma jituh , tau berjalan dengan labih bahalap .
Machine Translation Configuration 1 :
"Jadi, kita berharap dengan ada Bumdesma ini , bisa waktu dengan lebih baik .
Machine Translation Configuration 2 :
"Jadi, kita berharap dengan ada Bumdesma ini , bisa berjalan dengan lebih baik .
Machine Translation Configuration 3 :
"Jadi, kita berharap dengan adanya Bumdesma ini , bisa berjalan dengan lebih baik .
Target Sentence
"Jadi, kita berharap dengan adanya Bumdesma ini , dapat berjalan dengan baik lagi .

Case number-9
Input Sentence :
Wabup menambahkan, are produk tau dikelola awi Bumdesma tersewut ji sesuai dengan kebutuhan lewu te ' masing-masing huang rangka menggatang visi dan misi Bupati-Wakil Bupati Gumas, yakni smart agro, dan smart tourism.
Machine Translation Configuration 1 :
Wabup menambahkan, banyak produk bisa dikelola oleh Bumdesma tersebut yang sesuai dengan kebutuhan desa itu masing-masing dalam rangka mengusung visi dan misi Bupati-Wakil Bupati Gumas, yakni smart agro, dan smart tourism.
Machine Translation Configuration 2 :
Wabup menambahkan, banyak produk bisa dikelola oleh Bumdesma tersebut yang sesuai dengan kebutuhan desa itu masing-masing dalam rangka mendukung visi dan misi Bupati-Wakil Bupati Gumas, yakni smart agro, dan smart tourism.
Machine Translation Configuration 3 :
Wabup menambahkan, banyak produk bisa dikelola oleh Bumdesma tersebut yang sesuai dengan kebutuhan desa itu masing-masing dalam rangka mendukung visi dan misi Bupati-Wakil Bupati Gumas, yakni smart agro, dan smart tourism.
Target Sentence
Wabup menambahkan, berbagai produk bisa dikelola oleh Bumdesma tersebut yang sesuai dengan kebutuhan desa itu masing-masing dalam rangka

mendukung visi dan misi Bupati-Wakil Bupati Gumas, yakni smart agro, dan smart tourism.

Case number-10
Input Sentence :
Misalnya dunia pertanian huang rima ji luas, tau peternakan, jagung, parei.
Machine Translation Configuration 1 :
Misalnya dunia pertanian **dalam rima yang luas**, bisa peternakan, jagung, padi.
Machine Translation Configuration 2 :
Misalnya dunia pertanian **dalam arti yang luas**, bisa peternakan, jagung, padi.
Machine Translation Configuration 3 :
Misalnya dunia pertanian **dalam arti yang luas**, bisa peternakan, jagung, padi.
Target Sentence
Misalnya dunia pertanian **dalam arti luas**, bisa peternakan, jagung, padi.

3. The decreasing in the BLEU score for machine translation 3 against the translation machine 4 with a difference in monolingual corpus from 1164 lines of sentences become -0.14%, then from machine translation 4 against machine translation 5 with a difference in monolingual corpus from 770 lines of sentences become 392 lines of sentence had decreased against BLEU score -0.21% wherein the addition and subtraction of a monolingual corpus was less efficient in increase the result of the above translation 30%(Wahyuni et al., 2019) .

5. CONCLUSION
Based on the analysis and testing result, then the conclusions is statistical machine translation can be implemented to translate Dayak Language into Indonesia Language then he training corpus really affected the BLEU score, the more training corpus number, better the BLEU score but Monolingual corpus has not affected to the BLEU score significantly.

The best accuracy presentation value was found on machine 3 with the number of parallel corpus 806 sentence lines and 1164 monolingual which results in a BLEU score 49.15%.

6. LITERARY REFRENCE

Ansori, M. S. (2019). Sosiolingustik dalam kepunahan bahasa. *An-Nuha*, *6*(1), 52–61.

Asparilla, M. G., Sujaini, H., & Nyoto, R. D. (2018). *Perbaikan Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik (Studi Kasus : Bahasa Indonesia – Jawa Krama)*. *1*(2), 66–74.

Darwis, H. M. (2011). *The Fate of Regional Languages in the Era of Globalization: Opportunities and Challenges*. 1–13.

Dugonik, J., Bošković, B., Maučec, M. S., & Brest, J. (2015). The usage of differential evolution in a statistical machine translation. *IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - SDE 2014: 2014 IEEE Symposium on Differential Evolution, Proceedings*, *December*.

Hadi, I. (2014). Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia Ke Bahasa Melayu Sambas Dan Mesin Penerjemah Statistik (MPS) Bahasa Melayu Sambas Ke Bahasa Indonesia. *Jurnal Sistem Dan Teknologi Informasi*, *2*, 1–6.

Mandira, S., Sujaini, H., & Putra, A. B. (2016). Perbaikan Probabilitas Lexical Model Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *2*(1), 3–7. https://doi.org/10.26418/jp.v2i1.13393

Manual, U., & Guide, C. (2012). Statistical Machine Translation SystemUser Manual and Code Guide. *University of Edinburgh*, 1–267.

Mulyana, M., Sujaini, H., & Pratiwi, H. S. (2018). Algortima Pembagian Frasa Dalam Kalimat Untuk Menigkatkan Akurasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Bugis Wajo. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*, *6*(2), 39.

Sujaini, H. (2017). Meningkatkan Peran Model Bahasa dalam Mesin Penerjemah Statistik (Studi Kasus Bahasa Indonesia-Dayak Kanayatn). *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *3*(2), 51.

Wahyuni, M., Sujaini, H., & Muhardi, H. (2019). Pengaruh Kuantitas Korpus Monolingual Terhadap Akurasi Mesin Penerjemah Statistik. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*, *7*(1), 20.

Wentzel, G. (1922). Funkenlinien im Röntgenspektrum. *Annalen Der Physik*, *371*(23), 437–461.