JPPIPA 12(2) (2026)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

A Banjarnese Corpus Generation Method Based on Contextual Synonym Substitution Using Identic.v1.0 Data

Ali Muhammad[1*], Novia Winda[2], Budi Jejen Zaenal Abidin[3]

[1] Informatics Engineering Study Program, Faculty of Computer Science, Universitas Sains Indonesia, Bekasi, Indonesia.
[2] Indonesian Language and Literature Education, Faculty of Social and Humanities, Universitas PGRI Kalimantan, Banjarmasin, Indonesia.
[3] Information Systems Study Program, Faculty of Computer Science, Universitas Sains Indonesia, Bekasi, Indonesia.

Received: December 21, 2025
Revised: February 01, 2026
Accepted: February 25, 2026
Published: February 28, 2026

Corresponding Author:
Ali Muhammad
ali.muhammad@lecturer.sains.ac.id

DOI: 10.29303/jppipa.v12i2.14393

Abstract: The preservation and revitalization of the Banjar language is urgently needed. The decreasing number of Banjar language speakers and linguistic experts due to aging factors, combined with the hegemony of dominant languages brought by migrants, has become a major challenge in the preservation and revitalization of the Banjar language. This study aims to generate method for generating a Banjar language corpus by increasing the accuracy of sentence translation without leaving the original sentence context. This study uses a translation method of paraphrase contextual synonym substitution. This study used parallel corpus data Identic.v1.0. This method was tested and compared with statistical machine translation methods using Meteor universal tools, statistic evaluation and by human judgment. The statistical evaluation results indicate that the proposed method yielded a significant improvement in translation performance compared to the statistical machine translation method. Translation accuracy increased from 48% with the statistical method to 81% with the proposed method, representing a performance improvement of 33 percentage points, or approximately 68.75% relative to the statistical method. Meanwhile, the naturalness test of translated sentences using meteor universal tools with 1000 random sentences data shows that the proposed method is better than the previous method. The results or final score of naturalness sentences using proposed method are 0.6, while the final score of translating results using the statistical machine translation method is 0.36. Finally, the sentences evaluated by human judgment involving 15 language observers. The evaluated results show that the translated sentences using the proposed method is 75.8% more better than the statistical machine translation method.

Keywords: Contextual synonym substitution; Corpus generation methods; Human minimal resources; Translation methods.

Introduction

The Banjarese language shift is an issue that requires special attention. The phenomenon of language shift occurs when a community collectively no longer uses the language as its everyday language (Álvarez-carmona et al., 2022; Aqlan et al., 2019). This Banjarese language shift is caused by the hegemony of other regional languages, Indonesian, and international languages (Barmawi et al., 2023; Barmawi & Muhammad, 2019). In addition to language hegemony, immigrants and transmigrants have contributed to a 56.17% decline in the Banjarese population in 2022, exacerbating the Banjarese language shift (Fashwan & Alansary, 2021; Guinovart, 2019). This Banjarese language shift is characterized by a decline in the number of Banjarese speakers, both young and old. Furthermore, language shift can also be identified by the use of mixed languages in everyday speech (Hapsari et al., 2021). The brink of language extinction can be seen in the lack of young people understanding the meaning of Banjarese vocabulary (Fashwan & Alansary, 2021). A corpus is a collection of authentic texts selected and

How to Cite:
Muhammad, A., Winda, N., & Abidin, B. J. Z. (2026). A Banjarnese Corpus Generation Method Based on Contextual Synonym Substitution Using Identic.v1.0 Data. *Jurnal Penelitian Pendidikan IPA*, *12*(2), 487–498. https://doi.org/10.29303/jppipa.v12i2.14393

categorized for further language processing by computer machines. The benefit of creating a Banjarese language corpus is to facilitate computer linguistics experts in natural language processing of Banjarese (Ginting et al., 2025; Hasmianti et al., 2023). Such as automatic translation machines (Hasmianti et al., 2023; Kamariah et al., 2023), question and answer machines, summarization machines, computational linguistics (Fashwan & Alansary, 2021; Larasati, 2012; Liu & Huang, 2021; Lopez, 2023), sentiment analysis (Mohammed, 2022; Muhammad & Kamariah, 2020), linguistic steganography (Muhammad & Widyastuti, 2024; Muttaqin, 2019; Winda & Muhammad, 2023) and so on. Research in computer linguistics related to the Banjar language is carried out as an effort to conserve and revitalize it through language technology. The availability of a Banjar language corpus is very necessary for natural language processing, especially the Banjar language. Until now, there has been no Banjar language corpus that can be used as research material in the field of computer linguistics.

Previous research has attempted to translate sentences and even build language corpora using neural machine translation techniques such as Attention Mechanism, Large Language Model, Long Short-Term Memory, Subword Segmentation, and Syntactical Machine Translation (SMT) (Siu, 2023; Lyu et al., 2024; Team, 2024; Shah et al., 2023; Isnaeni et al., 2024; Wardhana et al., 2024; Tan et al., 2021; Saunders et al., 2020; Das et al., 2024; Dewangan et al., 2021). However, these approaches have been ineffective due to weaknesses in scalability, time efficiency, and less natural translation results. Meanwhile, synonym contextual substitution methods have been applied to majority languages and linguistic steganography, but have not been systematically adapted for regional languages with limited human resources (Barmawi & Muhammad, 2019; Muhammad & Widyastuti, 2024; (Muhammad et al., 2025; Winda & Muhammad, 2023).

Contextual synonym substitution is a paraphrasing method in natural language processing. Paraphrasing is a crucial component in various Natural Language Processing (NLP) applications, such as linguistic steganography, recommendation systems, and machine translation. One common approach is synonym substitution, where a word in the original sentence is replaced with another word with a similar meaning. Various previous studies have applied this technique, such as the DIRT method, which relies on syntactic relationship patterns; the bilingual pivoting approach, which utilizes bilingual translation data; and the PPDB method, which constructs a large database of paraphrase pairs from parallel data (Barmawi & Muhammad, 2019). While these methods have their respective advantages, their main weakness is that they do not consider the context in which words are used in a sentence in depth, often resulting in unnatural paraphrases. Another popular method is the N-gram-based method introduced by Gadag and Sagar, which relies on trigram similarities in a corpus to determine word substitutions (Gadag & Sagar, 2016). However, this approach still has low accuracy—around 46.3%—because it does not consider syntactic structure and semantic context, often resulting in sentences that are inconsistent with the original meaning. The advantage of the contextual synonym substitution method lies in the accuracy of sentence translation, ensuring that the semantic structure and context match. As a result, sentences can be fully translated without losing their original meaning. However, this process has drawbacks, one of which is the high computation time due to the processing steps required.

The limitations identified in previous research have prompted the need for a synonym substitution approach that takes context into account more comprehensively. The Contextual Synonym Substitution method introduced in this article offers a solution by combining N-gram-based probabilistic analysis with linguistic evaluation, including Part-of-Speech (POS) tagging and syntactic structure examination (Muhammad & Widyastuti, 2024) . This method performs a more selective word filtering process based on frequency of occurrence in the corpus, word class appropriateness, and synonym suitability through a probability score (NGM score). Thus, the synonyms used not only share meaning but also align with the sentence context. The next stage is syntactic structure evaluation to ensure that the paraphrased results retain the grammatical patterns and relationships of the original sentence (Barmawi & Muhammad, 2019). This approach has been proven to produce more natural paraphrases, based on testing using METEOR and human assessment, and even surpasses previous methods when tested on data from outside the training corpus.

Overall, the literature indicates that the development of paraphrasing methods has moved from a solely pattern-similarity-based approach to methods that are more contextual and sensitive to language structure. The main contribution of this contextual synonym substitution method is its success in overcoming the weaknesses of the traditional N-gram method and providing a more suitable approach for Indonesian, which has specific morphological and syntactic characteristics. By taking context into greater depth, this method offers significant improvements in naturalness and meaning preservation, making it a more effective solution in modern paraphrasing applications (Barmawi & Muhammad, 2019).

Previous corpus research has been conducted on professional electronic content (Nur et al., 2023), English

language learning (Oliver, 2024), tourist translation in the new media era (Pan & Qin, 2022; Prabowo & Indra Sanjaya, 2024). Methods in corpus formation research include using algorithm implementations based on ajax+jquery (Rui & Xiuli, 2022), grammar approaches (Shen, 2022), multimedia and lexical semantics (Spatioti et al., 2022) and morphological approaches with Identic data (Larasati, 2012) . This research aims to conserve and revitalize the Banjar language through the formation of a Banjar language corpus. To produce a parallel corpus of the Banjar language, this study uses the translation of contextual synonym substitution method (Liu & Huang, 2021) using Identic.v1.0 data (Larasati, 2012), the Indonesian thesaurus (Sudibyo, 2008), the Banjar language dictionary (Hapip, 2007), and the Palui sentence collection (Team, 2025). The Banjar language corpus was then tested using the METEOR (Metric for Evaluation of Translation with Explicit ORdering) testing method to ensure that the Banjar language corpus has met the standards as a parallel corpus. The availability of a Banjar language corpus as a source of training data in machine learning-based applications or computational linguistics is very important (Muhammad et al., 2025; Winda & Muhammad, 2023). This research aims to build a corpus of regional languages (Banjar language) in the face of a scarcity of experts. This research will be a contribution of Banjar Language and Literature Science to natural language processing which is useful for regional, national, and international interests. In addition, the existence of a Banjar language corpus is also an effort to conserve and revitalize regional languages, especially if language experts are difficult to find.

Method

The development of the Banjar language corpus in the proposed research uses the contextual synonym substitution method. This Banjar language corpus was developed using the Python programming language and did not use any specific libraries (such as Sastrawi) during development, either for the tagging process or during the evaluation and selection of translated sentences. The entire contextual synonym substitution method is explained in Figure 1 below.

The contextual synonym substitution method supported by parallel corpus data Identic.v1.0 begins with the preparation of research data in the form of a Banjar dictionary, an Indonesian thesaurus, and the Identic.v1.0 corpus. Identic.v1.0 corpus contains 29,934 sentences taken from various sources. Next, a word selection process is carried out from this data. The result or output of the word selection process is a filtered Indonesian thesaurus. The next process stage is translation of the Identic.v1.0 corpus using an application/program with data sources taken from a filtered Indonesian thesaurus, Identic.v1.0 corpus, and a Banjar language dictionary. The result of the Identic.v1.0 corpus translation process is a candidate Banjar language corpus. Finally, to generate a Banjar language corpus that is verified and structured according to Banjar language rules, a sentence structure evaluation process is carried out.

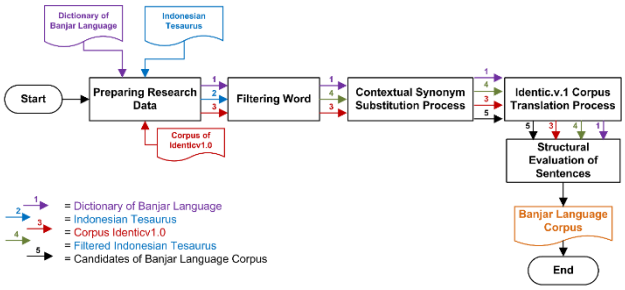


Figure 1. Proposed Method

1. *Preparing Research Data*

This study uses research data which include a Banjar language dictionary, a collection of Banjar language sentences, an Indonesian thesaurus, and the Identic.v1.0 corpus. The research data used are digital data. The list of Banjar language sentences were taken by crawling from mass media websites in the sipalui column (Team, 2025). The si Palui column is a fictional story board column writed in Banjar language and posted on Banjarmasin Post mass media. While the Banjar Language Dictionary already in the digital file, but needs preprocessed to be used. The preprocessing steps are explained in Figure 2.

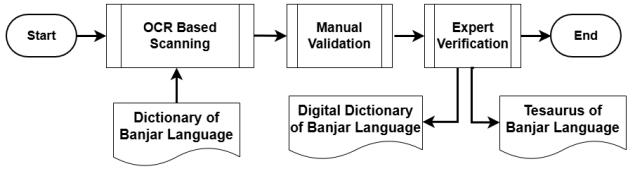


Figure 2. Preprocessing of Banjar Language Dictionary

The Banjar dictionary was initially a PDF file manually scanned from the Banjar dictionary and then scanned using OCR. The result of this scan was a Banjar dictionary file. Due to the very high level of OCR accuracy and the poor quality of the printed Banjar dictionary, there were errors in identifying words in the scanned results using OCR. Previous validation had been attempted using the naive Bayes method, but the results of processing using naive Bayes were not optimal. Manual validation was used to check and ensure that each word in the Banjar dictionary file was in accordance with the printed and digital versions of the Banjar dictionary. This manual validation process resulted in two files: a digital Banjar dictionary file and

a Banjar thesaurus. The Banjar thesaurus is a list of Indonesian words and synonyms words in Banjar language. Finally, all files were verified by experts to ensure that the Banjar dictionary file and the Banjar thesaurus file is appropriate with Banjarese grammar.

2. *Filtering Word Process*

Filtering word process is proposed during the preprocessing of the Indonesian thesaurus. This is used to ensure that during the transformation and translation process, the Indonesian thesaurus words into Banjarese have their respective word pairs. The Filtering word process process is explained in Figure 3.

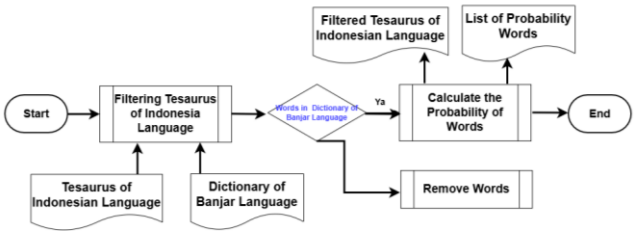


Figure 3. Filtering Word Process

The filtering word process begins with the Indonesian thesaurus filtering process. This Indonesian filtering process has two inputs, namely the Indonesian thesaurus and the Banjar dictionary file. If Indonesian thesaurus words contained in the Banjar dictionary, the word conducted a word probability calculation process. From the list of word probability, a banjarese thesaurus words will be generated with their occurrence probabilities. These words and probabilities are useful for graph formation during the sentence translation process based on contextual synonym substitution. Furthermore, Indonesian thesaurus words that are not in the Banjar dictionary will be removed.

3. *Contextual Synonym Substitution Process*

Contextual synonym substitution process is proposed for substituting words in the corpus with their synonyms. There are two processes at this stage, which can be seen in Figure 4.

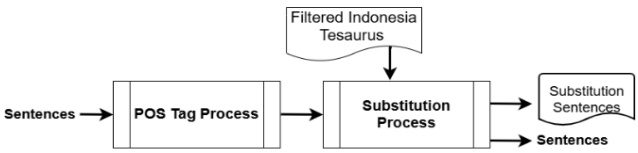


Figure 4. Contextual Synonym Substitution Process

i. POS Tag Process

POS tagging process is used to assign word classes to each word based on its syntactic function (POS: part of speech), such as noun, verb, adverb, adjective, and so on. For POS tagging, a hidden Markov model is used (Muhammad & Widyastuti, 2024). POS tagging involves tokenizing words in an Indonesian dictionary. This method uses state transition diagrams and probabilistic (rule-based) methods to determine the appropriate tag for a word.

These words are divided into two classes: closed-class words and open-class words. Open-class words are words such as nouns, verbs, and adjectives, while closed-class words are words such as pronouns and conjunctions. The POS tagging process begins with processing closed-class words, followed by processing open-class words. When ambiguity occurs, predefined rules are used to predict or find the appropriate word tag.

Example of the tagging process for Indonesian: apakah kamu sedang meyindir ali ←→ apakah/**PRON** kamu/**PRP** sedang/**RB** Menyindir//**VBT** ali/**NN.**

ii. Substitution Process

The synonym substitution method was conducted in this process based with interchangeable words, namely words that can be exchanged with each other, even for different meanings. To find interchangeable words, a set of synonyms is generated based on the occurrence of words in the corpus by calculating the word occurrence probability (Barmawi & Muhammad, 2019). The frequency of occurrence (fn) of the word 'menyindir' in the corpus is shown in Table 4. In this case, for fn is equal to 2-gram, 3-gram, and 4-gram, 5-gram. Then, the count function (Count(w)), (where w is the word 'menyindir') is calculated using Equation (1).

$$\text{Count(w)} = \sum_2^m \log f_n \quad (1)$$

Table 1. Frequency of n-gram

n-gram	Frequency	fn
sedang menyindir	2	4
menyindir hana	1	
menyindir siapa	1	
kamu sedang menyindir	2	4
sedang menyindir hana	1	
sedang menyindir siapa	1	
apakah kamu sedang menyindir	1	4
hei kamu sedang menyindir	1	
kamu sedang menyindir hana	1	
kamu sedang menyindir siapa	1	

Next, the maximum value of Count(w) (called $\max_{count}$) is obtained, followed by calculating the proportion between the 'menyindir' count and $\max_{count}$. This proportion is called NGM_{score}(menyindir), as shown in Equation (2).

$$Score_{NGM}(w) = \frac{\text{count(w)}}{max_{count}} \quad (2)$$

NGM_{score} used in the synonym generation process. To generate a set of synonyms (synsets), a threshold of NGM_{score} is determined by probability of original word. Words with a NGM_{score} greater than or equal to the threshold are included in the synset graph, otherwise, they are removed from the graph. The same process is applied to the words 'kamu', 'sedang', 'menyindir'. Examples of synonym substitution in Indonesian are shown in Figures 5 (a), (b), and (c). Based on Figures 5 (a), (b), and (c), the word 'kamu' has three synsets. The first synset is {tuan}, the second is {situ}, and the third is {kau, awak, kakak}. The word 'sedang' has three synsets. The first synset is {lagi, masih, tengah}. The second synset is {selagi, sementara}, and the third synset is {pantas, mampu, cukup, pas, cukup}. Lastly, the word 'menyindir' has 3 synsets including the first synset {mengusik}, the second synset {ejek, merendahkan}, the third synset {menyakiti, menodai, mengejek, mengucilkan, memaki, melecehkan, hina}. The value that follows the word is the NGM_{score} value. Suppose the threshold applied in the synset 'kamu' is 0.2, in the synset 'sedang' is 0.53 and 0.15 in the synset 'menyindir'. All words in the synset 'kamu' with an $NGM_{score} \leq 0.2$ are removed. Likewise, words in the synset 'sedang', words with an $NGM_{score} \leq 0.53$ are removed and all words in the synset 'menyindir' with an $NGM_{score} \leq 0.15$ are removed.

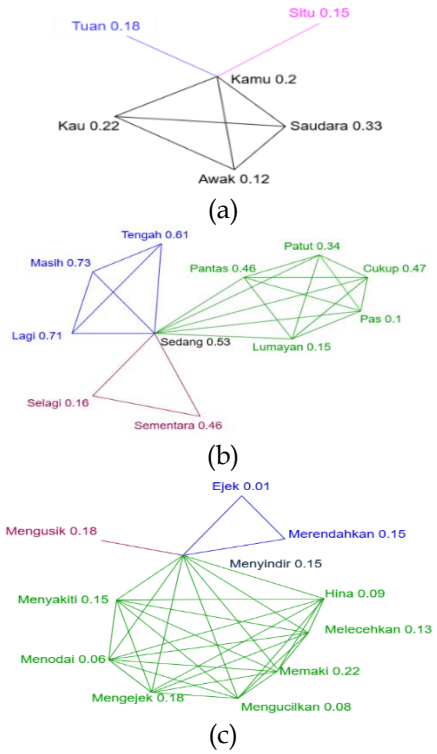


Figure 5. Synonyms of graph in the word, (a) 'kamu', (b) 'sedang', (c) 'menyindir'

The synonym graphs after conducted the threshold using the probabilies of original words, unrequirements word are removed. The example of this process shown in Figures 6 (a), (b) and (c). The Words with high NGM_{score} have a high probability of being translated candidates and have a high opportunity of being a natural translation. For example, the word 'kau' can be used to replace the word 'kamu', the word 'tengah' can be used to replace the word 'sedang', and the word 'mengejek' can be used to replace the word 'menyindir'.

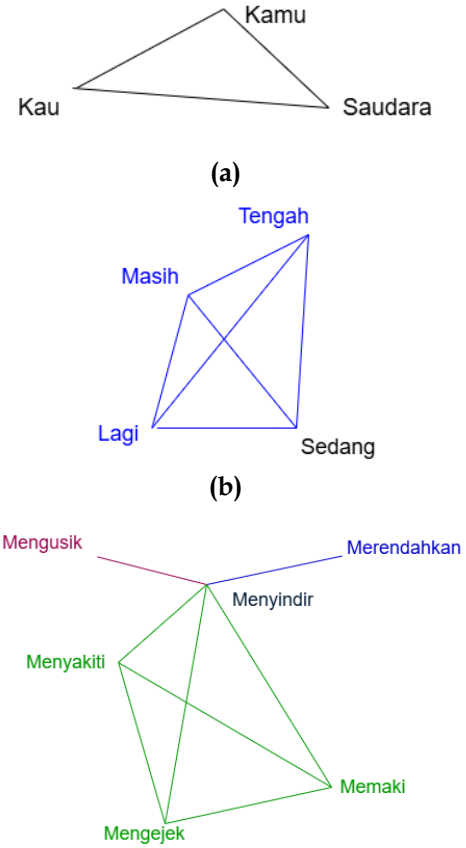


Figure 6. Synonym Graph of Words (a) 'kamu', (b) 'sedang', (c) 'menyindir', after word removed

After applying word substitution, the paraphrased sentences are grouped into a list of substitution sentences. The list of substitution sentences from this process is shown in Table 2.

Table 2. List of substitution sentences

List of Substitution Sentences
apakah kamu sedang menyindir hana
apakah kau sedang menyindir hana
apakah saudara sedang menyindir hana
⋮
apakah kamu sedang mengejek hana

4. Translation of the Identic.v1.0 Corpus Process

Translation of the Identic.v1.0 Corpus process was conducted to translate candidate sentences in the substitution sentence list into Banjarese sentences. This translation process involves replacing each word from the substitution sentence list with words in the Banjarese dictionary. This sentence translation process is explained in Figure 7.

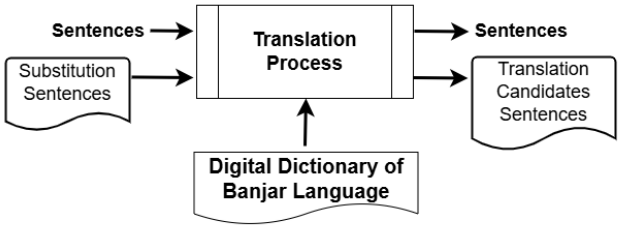


Figure 7. Translation Process

Each word in the list of substitution sentences was translated into Banjarese. This was done to generate a translation that similar with the original sentences, both grammatically and contextually. The results of the corpus translation process are shown in Table 3.

Table 3. Translation Process Results

List of Substitution Sentences	Translation in Banjar Language
apakah kamu sedang menyindir hana	apakah ikam sadang menyindir hana
apakah kau sedang menyindir hana	apakah ikam sadang menyindir hana
apakah saudara sedang menyindir hana	apakah ikam sadang menyindir hana
⋮	⋮
apakah kamu sedang mengejek hana	apakah ikam sadang mauha-rakan hana

5. Structural Evaluation Process

Structural evaluation is a method for filtering a list of contextual synonym substitution sentence translations based on the sequence tags of the original sentences. This process is divided into two sub-processes: selecting maximum Banjar words and sentence structure evaluating process. This process requires two input data: the sentencec and a list of translation candidate sentences. The details of this process are outlined in Figure 8.

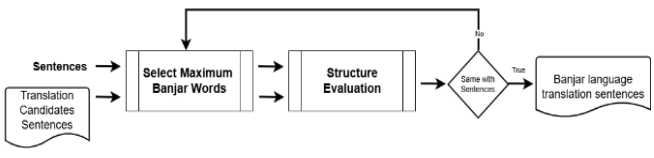


Figure 8. Structural Evaluation Process of Sentences

i. Selecting Maximum Banjar Words Process

The process of selecting maximum Banjar words is proposed to maximize the number of translated words into Banjar sentences. The main reason for this process is the lack of a Banjar thesaurus. Banjar words sometimes have specific translations for one word and do not consider similarities with other words. This process allows to generate and select only the maximum number of translated words in Banjar. Each sentence in the list of candidate translations will be counted. The sentence with the largest number of translated words will be used as input for the structure evaluation process. The output of this process can be seen in Table 4.

Table 4. Number or Translated Words

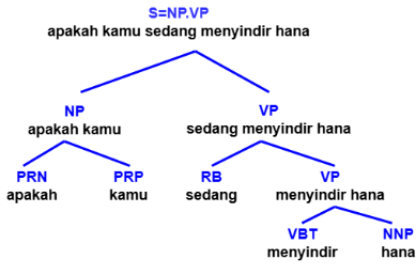
Translation Sentence in Banjar Language	Number of translated words
apakah ikam sadang menyindir hana	2
apakah ikam sadang menyindir hana	2
apakah ikam sadang menyindir hana	2
⋮	⋮
apakah ikam sadang mauharakan hana	3

In Table 3 above, each sentence will have its own number of translated words. The sentence with the largest number of translated words, "apakah ikam sadang haurakan hana" is selected anda used as input for the sentence structure evaluation process in the next process.

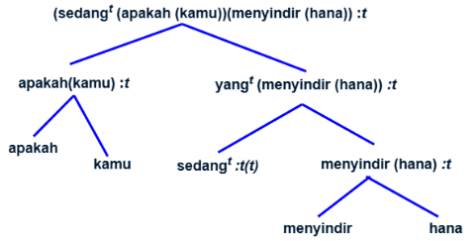
ii. Sentence Structure Evaluation Process

The sentence structure evaluation process proposed for producing contextually appropriate translations based on the sentence's syntactic structure. In this case, if the sentence sequence tags in the substitution sentence translation list are required to match the word sequence tags in the original sentence, then the sentence in the candidate sentence list can be used as the translated sentence.

For example, if we have the original sentence "apakah kamu sedang menyindir hana", the syntactic structure and semantic interpretation are shown in Figure 9. The results of POS tagging process is "apakah/**PRN** kamu/**PRP** sedang/**RB** menyindir/**VBT** hana/**NN**".



(a)



(b)

Figure 9. (a) syntactic structure and (b) semantic interpretation of 'apakah kamu sedang menyindir hana'

The word 'kamu' is marked as a personal pronoun in the synset. So it can be translated into the word 'ikam'. Word "kamu" and "ikam" has the same tag in the original sentence. Likewise, the word 'sedang' is marked as an adverb in the synset so it can be translated into the word 'sadang' because that words has the same tag in the original sentence. Finally, the word 'menyindir' is marked as a transitive verb in the synset so it can be translated into the word 'mauharakan' because it has the same tag in the original sentence. Finally, the word 'ikam' is chosen as the translation of 'kamu', the word 'sadang' is chosen to be the translation of 'sedang', and the word 'mauharakan' is chosen to be the translation of 'menyindir'. The selected translation result is 'apakah ikam sadang mauharakan hana'. The syntactic structure and semantic interpretation of the translated sentence are shown in Figure 10.

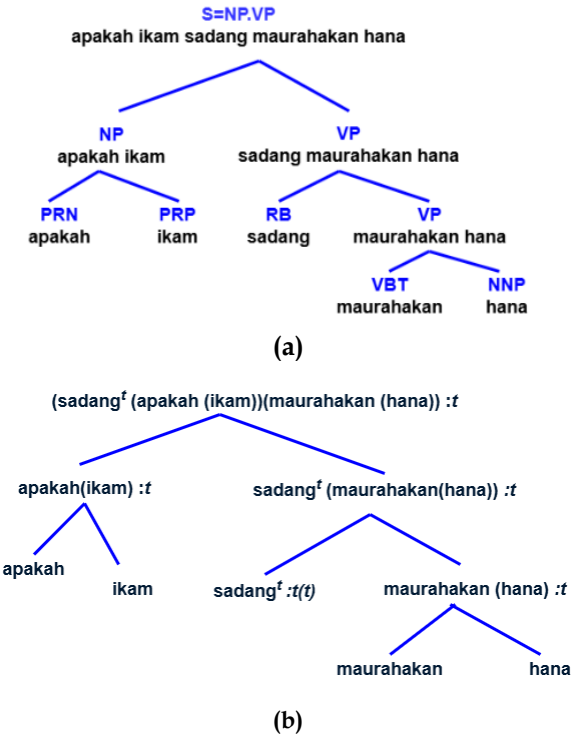


Figure 10. (a) syntactic structure and (b) semantic interpretation of translated sentences

Result and Discussion

This section describes the methods and tools for evaluation and the results of the experiments.

1. Naturalness Evaluation

This study uses the Identic.v1.0 corpus, and the corpus from the 'si palui' column from the Banjarmasin Post newspaper (Team, 2025). The purpose of selecting the 'si palui' column is to obtain formal and natural Banjarese sentences. To obtain formal and natural sentences, regular expressions such as 'hashtag', 'username', etc. are removed. Two methods are used to evaluate the naturalness of translated sentences: metric evaluation and human assessment. This study uses Identic.v1.0 data and the si palui corpus to test the naturalness of the translated sentences. The evaluation is carried out using metric evaluation (meteor) (Banerjee & Lavie, 2005).

i. Metric Evaluation (Meteor Universal Tools)

To evaluate the performance of the proposed method, an n-gram-based evaluation is required. N-gram-based evaluation is performed by applying a penalty. Meteor evaluates translated sentences by calculating a score based on word-for-word matches between the translated sentence and the reference sentence.

The procedure for assessing the reasonableness of a translated sentence is as follows:

1. Write a list of all possible unigram mappings from the translated word to the reference sentence.
2. Select the largest list of unigram mappings, such that each unigram in the translated sentence can only be mapped to one word in the reference sentence.
3. Calculate the precision, which represents the system's accuracy. System accuracy only considers the number of matched unigrams. The precision calculation is formulated as follows:

$$\boldsymbol{P} = \frac{\sum_i \boldsymbol{w_i}.(\boldsymbol{\delta}.\boldsymbol{m_i}(\boldsymbol{h_c})+(\boldsymbol{1}-\boldsymbol{\delta}).\boldsymbol{m_i}(\boldsymbol{h_f}))}{\boldsymbol{\delta}.|\boldsymbol{h_c}|+(\boldsymbol{1}-\boldsymbol{\delta}).|\boldsymbol{h_f}|} \quad (3)$$

where P is precision, $\sum i$ is the total number of test words, w_i is the observation word, m_i is the number of matched translated words, h_c is the content word, dan h_f is the function word covered by the matched word in the test sentence, and δ is 10^{-3}.

4. The recall calculation indicates the system's accuracy in finding word fragments from translated sentences that appear in test sentences. The system's accuracy is based solely on the number of matched unigrams. The recall calculation is formulated as follows:

$$\boldsymbol{R} = \frac{\sum_i \boldsymbol{w_i}.(\boldsymbol{\delta}.\boldsymbol{m_i}(\boldsymbol{r_c})+(\boldsymbol{1}-\boldsymbol{\delta}).\boldsymbol{m_i}(\boldsymbol{r_f}))}{\boldsymbol{\delta}.|\boldsymbol{r_c}|+(\boldsymbol{1}-\boldsymbol{\delta}).|\boldsymbol{r_f}|} \quad (4)$$

where R is recall, $\sum i$ is the total number of test words, w_i is an observation word, m_i is the number of matched test words, r_c *is the word content and* r_f is a function word that is covered by a matching word in the translated sentences, and δ is 10^{-3}.

5. Calculating the aggregate precision and recall scores, the harmonic mean F1 is:

$$\boldsymbol{F_1 = \frac{2.P.R}{P+R}} \quad (5)$$

where P is Precission and R is Recall

6. The translated acceptability evaluation is performed by measuring the similarity of the semantic framework and its role fillers between the reference and translated sentences, which is represented by F_mean. Meteor is often considered a recall-oriented metric; this metric uses alpha (α) as a control for the relative weight between precision and recall. The specified alpha value is 0.9 so that the F_mean result is concluded as natural language and in accordance with human judgment perception. F_mean is calculated as follows:

$$\boldsymbol{F_{mean} = \frac{P.R}{\alpha.P+(1-\alpha).R}} \quad (6)$$

where P is precission, R is Recall, and α = 0.9 (Banerjee & Lavie, 2005).

7. To evaluate a translation based on n-grams, it is necessary to measure the naturalness of the translated sentence and the correlation between the translated sentence and the reference sentence, which appear to be the same or have the same meaning. The closeness of the sentence's meaning is concluded based on the results of a comparison between the smallest chunks of the translated sentence and the reference sentence. Chunks are defined as adjacent and identical sequences between words in the test sentence and words in the translated sentence. Suppose we have a test sentence '**apakah kamu sedang menyindir hana'**, and we have the translated sentence ***'apakah ikam sadang mauharakan hana'***. Then, we have to find the same word sequence '***apakah ikam sadang meuharakan hana***' in the translated sentence. Word order '***apakah ikam sadang meuharakan hana***' cut from the test sentence are called fragments/chunks. The fragmentation penalty is calculated as the number of deductions divided by the number of matching candidate words. Suppose ϒ set a maximum penalty and β establishes a functional relationship between fragmentation and penalty. The fragmentation penalty (Pen) is calculated as follows:

$$\boldsymbol{Pen = \Upsilon(\frac{c_h}{m})^{\beta}} \quad (7)$$

where c_h is the number of test sentence fragments and m is the number of matching unigrams, ϒ is 0,5 and β is 3,0 (Banerjee & Lavie, 2005).

8. Finally, the final score calculation to represent the total aggregate value consists of precision, recall, F_{mean} , as shown in the following equation:

$$FinalScore = (1 - Pen).F_{mean} \quad (8)$$

ii. Human Judgment Evaluation

This evaluation relies on human perception, which is related to knowledge and experience. Several methods exist for evaluating human judgment, such as interviews, questionnaires, and polls. This study used 15 respondents divided into two groups, namely 10 linguist or expert respondents and 5 non-expert respondents. These experts are Banjar language researchers, Banjar language observers, journalists and Banjar language teachers. Non-expert respondents are those who actively access, read, and write newspapers, such as students, lecturers, researchers, and the general public. Each respondent must evaluate 15 translated sentences using the statistical machine translation method compared to 15 translated sentences using the proposed method with the original sentences. Respondents assessed the naturalness of the translation results using percentages based on linguistic aspects of the Banjar language such as morphology, sentence structure, grammar, and politeness. Respondents must determine the unnaturalness of the sentences and comment on incorrect words. Naturalness is represented by the percentage of naturalness using Equation (9):

$$Naturalness\ Percentage = \frac{Nat_{trans}}{Nat_{ori}} * 100\% \quad (9)$$

where Nat_{trans} is the naturalness of the translated sentence and Nat_{ori} is the naturalness of the original sentence.

2. Discussion

This section discusses and analyzes the results of Meteor and human assessment.

i. Meteor Universal Tools Results

This study used samples of 10, 50, 100, 500, and 1000 sentences from the identical.v.1 corpus, translated sentences using the statistical machine translation method, translated sentences using the contextual synonym substitution method from the si palui corpus. The test results are divided into two, namely the Percentage Results of Translated Sentences and Translation Metric Evaluation Results.

a. Percentage Results of Translated Sentences

The percentage of translated sentences was analyzed to determine the maximum effectiveness of the contextual synonym substitution method for translating the corpus from Indonesian to Banjar. This method of testing translated sentences was compared with a statistical-based translation method. This evaluation involved 10, 50, 100, 500, and 1,000 sentences taken from the original sentences of the identical.v.1 corpus. The

translated sentences were translated using the statistical machine translation method and the translated sentences were translated using the contextual synonym substitution method. The evaluation results using statistic can be seen in the graph below, Figure 11.

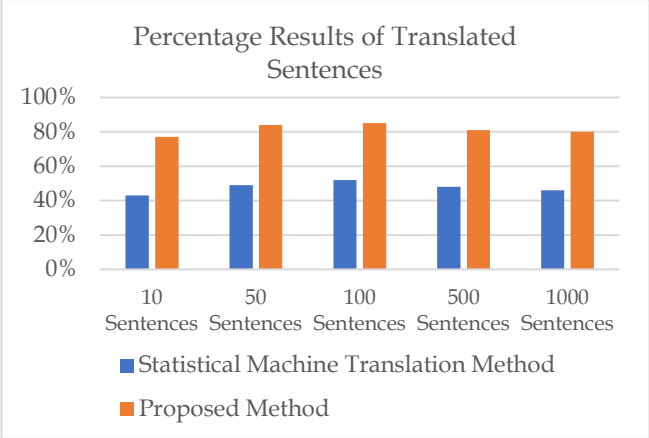


Figure 11. Graph of the results of the percentage of translated sentences

The graph 11 above shows the results of a comparative test of the percentage of translated sentences taken randomly from the Identic.v1.0 corpus. The sentences were then translated using the statistical machine translation method and translated using the contextual synonym substitution method. The test results show that the percentage of using the contextual synonym substitution method is more effective than using the statistical machine translation method. This can be seen from the graph above, for 10 original sentences, 77% can be translated using the contextual synonym substitution method. Meanwhile, the statistical machine translation method can only translate 43% of the sentences. In the use of the next 100 test sentences, the contextual synonym substitution method can translate 85% of the sentences, while the statistical machine translation method is only able to translate 52%. The average percentage of translation results using the contextual synonym substitution method overall was 81%, while the average percentage of translation results using the statistical machine translation method was 48%. This shows that the use of the contextual synonym substitution method is more effective than using the statistical machine translation method with an increase in the percentage of translation success of 68.75%. This percentage increase was obtained from the calculation of translation results using the statistical method of 48% to the proposed method, which is 81%. This percentage represents a performance increase of 33 percentage points, or approximately 68.75% relative to the statistical method.

b. Translation Metric Evaluation Results

Translation metric evaluation is used to evaluate machine translation that incorporates linguistic features such as synonyms, stemming, and word order, and places greater emphasis on recall to better align with human assessment of translation quality. To test the performance of the proposed method, this translation metric evaluation involves 10, 50, 100, 500, and 1,000 sentences taken from various original translated sentences from the Identic.v1.0 corpus, the Si Palui corpus, translated sentences using the statistical machine translation method, and translated sentences using the contextual synonym substitution method. Furthermore, the naturalness of the sentences produced by the statistical machine translation method and the proposed method is compared. The results of the translation evaluation test using the Meteor Universal Tools can be seen in the graph in Figure 12.

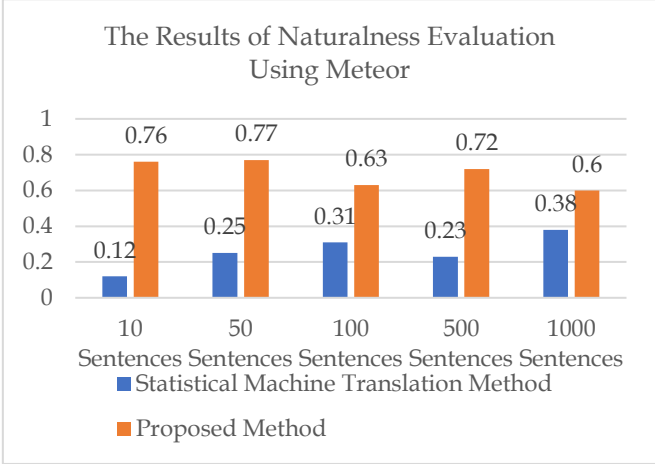


Figure 12. Results of translation evaluation using Meteor Universal Tools

Figure 12 above shows a comparison of translation evaluation results using the Meteor Universal Tools between the statistical machine translation method and the contextual synonym substitution method. The final result or score of sentence naturalness using 1000 sentences for the proposed method is 0.6, while the final score of translation results using the statistical machine translation method is 0.36. The evaluation results of the contextual synonym substitution method showed better results than the statistical machine translation method. This occurs because the statistical machine translation method does not consider the involvement of synonyms for the words to be translated, even though both methods maintain the grammatical structure of the original sentence. Both methods use changes during the translation process based on n-grams applied to each original sentence. The statistical machine translation method is only able to translate sentences based on the

words and the highest probability of the available words without changing words based on the list of synonyms, taking into account the context of each sentence. This makes the naturalness of translated sentences using the proposed method higher than the statistical machine translation method.

ii. Experimental Results Using Human Judgment

The results of the experiment using human judgment, as shown in Figure 13, show that the naturalness of sentence translation using the contextual synonym substitution method is better than using the statistical machine translation method. The reason is the same as explained in subsection 4.2.1.2 above.

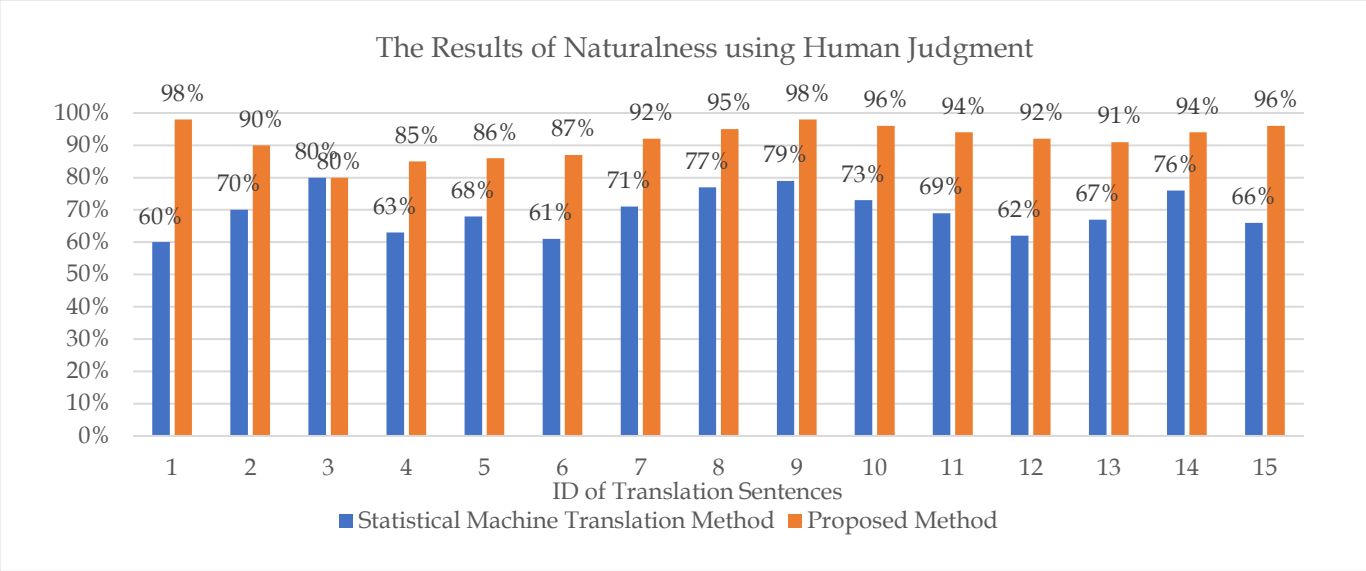


Figure 13. Results of the Translation Naturalness Experiment Using Human Judgment

Figure 13 above shows the results of the naturalness experiment of sentence translations using human judgment. The results of human judgment indicate that the translation results using the contextual synonym substitution method tend to be better than those using the statistical machine translation method. However, there are translated sentences that have the same naturalness between the two methods. From the figure above, it can be seen that the translation results of sentences in experiment ID 3 have the same level of naturalness. This naturalness occurs because there are no translated words that can represent each translation method, for example, the words in the sentence are not from words that exist in Indonesian. In this case, the context of the translated sentences using the statistical machine translation method and the proposed method are the same. Therefore, the naturalness of sentences based on human assessment is side by side. The evaluated results also show that the translated sentences using the proposed method is 75.8% more better than the statistical machine translation method.

Conclusion

The limitation of this study is the use of the contextual synonym substitution method to build the Banjar language corpus. In addition, reliance on manual dictionaries impacted the length of the research process. However, this became a challenge for the success of this study. Based on the evaluation results of human assessment and the Meteor Universal Tools, the sentence translation method using the contextual synonym substitution method produces better sentence translations than the statistical machine translation method. This contextual synonym substitution method has been proven to be applicable to overcome the problem of the lack of language experts during the formation of a language sentence corpus. The formation of a language corpus is a vital effort in language conservation activities considering that Indonesia has 1,340 diverse ethnic groups spread across Indonesia. In some translation cases, the naturalness of translated sentences can be similar between the statistical machine translation method and the contextual synonym substitution method. This occurs because there is no ambiguity or equivalent translated words. The insignificant difference in naturalness between the sentences translated using the statistical machine translation method and the contextual synonym substitution method is a contribution of the proposed sentence translation method. This study can be used as a

reference in the formation of other language corpuses if linguists are increasingly scarce and difficult to find.

Acknowledgments
The authors would like to express their gratitude to the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia (Kemendikbudristek). We would also like to express our gratitude to the academic community of Universitas Sains Indonesia for their institutional support as a research location so that this research can be completed well.

Author Contributions
This work was a collaborative effort among three authors: Ali Muhammad as the lead author, overseeing research conceptualization, data analysis, and initial manuscript preparation, while both Novia Winda and Budi Jejen Zaenal Abidin contributed equally by providing critical input on methodology, interpreting findings, and refining the final manuscript. All authors participated actively in discussions and approved the submitted version.

Funding
This research was funded by the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia (Kemendikbudristek) through the 2025 Beginner Lecturer Research Grant Program (PDP).

Conflicts of Interest
The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

References

Álvarez-carmona, M. Á., Aranda, R., Rodríguez-gonzalez, A. Y., Fajardo-delgado, D., Guadalupe, M., Pérez-espinosa, H., Martínez-miranda, J., Guerrero-rodríguez, R., Bustio-martínez, L., & Díaz-pacheco, Á. (2022). Natural language processing applied to tourism research : A systematic review and future research directions. *Journal of King Saud University – Computer and Information Sciences Xxx*, *xxx*(xxxx), xxx. https://doi.org/10.1016/j.jksuci.2022.10.010

Aqlan, A. A. Q., Manjula, B., & Naik, R. L. (2019). A Study of Sentiment Analysis : Concepts , Techniques , and Challenges. *Proceedings OfInternational Conference on Computational Intelligence and Data Engineering, Lecture Notes on Data Engineering and Communications Technologies 28*, 147–162. https://doi.org/10.1007/978-981-13-6459-4

Banerjee, S., & Lavie, A. (2005). METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Barmawi, A. M., & Muhammad, A. (2019). Paraphrasing method based on contextual synonym substitution. *Journal of ICT Research and Applications*, *13*(3), 257–282. https://doi.org/10.5614/itbj.ict.res.appl.2019.13.3.6

Barmawi, A. M., Wahyudi, B. A., & Pristi, T. (2023). Linguistic Based One Time Password. *International Journal on Electrical Engineering and Informatics -*, *15*(1), 1–16. https://doi.org/10.15676/ijeei.2023.15.1.1

Fashwan, A., & Alansary, S. (2021). A Morphologically Annotated Corpus and a Morphological. *Procedia Computer Science 189*, 203–210. https://doi.org/10.1016/j.procs.2021.05.084

Gadag, A. I., & Sagar, B. M. (2016). N-gram Based Paraphrase Generator from Large Text Document. *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 91–94.

Ginting, N. D. B., Sinaga, L. A., Ginting, A. S. B., & Surip, M. (2025). Pergeseran Bahasa Indonesia di Kalangan Remaja di Era Globalisasi. *Jurnal Multidisiplin Inovatif*, *9*(3), 124–129.

Guinovart, X. G. (2019). Enriching parallel corpora with multimedia and lexical semantics from the CLUVI Corpus to WordNet and SemCor. *John Benjamins Publishing Company*, 141–158. https://doi.org/https://doi.org/10.1075/scl.90.09gom

Hapip, A. D. (2007). *Kamus Banjar – Indonesia*. CV. Rahmat hafiz Al Mubaraq.

Hapsari, W. P., Labib, U. A., Haryanto, H., & Safitri, D. W. (2021). A Literature Review of Human, Organization, Technology (HOT) – Fit Evaluation Model. *Proceedings of the 6th International Seminar on Science Education (ISSE 2020), Advances in Social Science, Education and Humanities Research*, *541*(Isse 2020), 876–883. https://doi.org/10.2991/assehr.k.210326.126

Hasmianti, L., Usman, U., & Amir, J. (2023). Pergeseran Penggunaan Kata Sapaan oleh Generasi Milenial Banjar di Kota Banjarmasin. *Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, *8*(2), 122. https://doi.org/10.26737/jp-bsi.v8i2.4280

Kamariah, Hamidah, J., & Krismanti, N. (2023). Konservasi Bahasa Banjar Sebagai Usaha Pelestarian Bahasa Daerah di Kalimantan Selatan. *Bahasa, Sastra & Pengajaran (Konfiks)*, *10*(2), 24. https://journal.unismuh.ac.id/index.php/konfiksPermalink/DOI:https://doi.org/10.26618/jk/13118

Larasati, S. D. (2012). IDENTIC corpus: Morphologically enriched Indonesian-english parallel corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 902–

906.

Liu, B., & Huang, L. (2021). ParaMed : a parallel corpus for English – Chinese translation in the biomedical domain. *BMC Medical Informatics and Decision Making*, 1–11. https://doi.org/10.1186/s12911-021-01621-8

Lopez, A. (2023). Machine Translation evaluation metrics benchmarking : From traditional MT to LLMs. In *Universitat De Barcelona Fundamental* (1st ed.). Facultat de Matemàtiques i Informàtica, Universitat De Barcelona.

Mohammed, T. A. S. (2022). The Use of Corpora in Translation Into the Second Language: A Project-Based Approach. *Frontiers in Education*, *7*(April), 1–14. https://doi.org/10.3389/feduc.2022.849056

Muhammad, A., & Kamariah, K. (2020). Pengurai Kalimat Bahasa Banjar Dengan Menggunakan Parser PC-PATR. *Jurnal Linguistik Komputasional (JLK)*, *3*(1), 20. https://doi.org/10.26418/jlk.v3i1.30

Muhammad, A., & Widyastuti, N. (2024). Pengembangan Aplikasi Part-of-Speech Tagger Bahasa Banjar Menggunakan Metode Pengembangan DevOps. *JIKOMTI: Jurnal Ilmiah Ilmu Komputer Dan Teknologi Informasi*, *1*(1).

Muhammad, A., Winda, N., Firizkiansah, A., Setiawan, D., Dewi, S. H. F., Rizki, I. M., & Ardiansyah, M. (2025). Review of Banjarnese Neural Machine Translation Development With Minimal Resources. *Journal of Software Engineering, Information and Communication Technology (SEICT) 6(1)*, *6(1)*(June), 33–42. https://doi.org/https://doi.org/10.17509/seict.v6i1.86768

Muttaqin, A. I. (2019). Konstruksi Verba Gerak Direksional dalam Bahasa Banjar. *PRASASTI: Journal of Linguistics*, *4*(2), 99–103. https://jurnal.uns.ac.id/pjl/article/view/34129

Nur, S., Assyifa, A. N., & Nurjannah, H. (2023). Pengembangan Aplikasi Penerjemah Bahasa Isyarat Indonesia (Bisindo) Menggunakan Metode Long-Short Term Memory. *EDUSAINTEK: Jurnal Pendidikan, Sains Dan Teknologi*, *11*(1), 13–30. https://doi.org/10.47668/edusaintek.v11i1.898

Oliver, A. (2024). LitPC : A set of tools for building parallel corpora from literary works. *Proceedings Ofthe 1st Workshop on Creative-Text Translation and Technology, European Association for Machine Translation*, 21–31.

Pan, B., & Qin, Q. (2022). Construction of parallel corpus for english translation teaching based on computer aided translation software. *Computer-Aided Design and Applications*, *19*(s1), 70–80. https://doi.org/10.14733/CADAPS.2022.S1.70-80

Prabowo, A., & Indra Sanjaya, F. (2024). Penerapan Metode Transfer Learning Pada Indobert Untuk Analisis Sentimen Teks Bahasa Jawa Ngoko Lugu. *Jurnal Sistem Informasi Dan Sistem Komputer*, *9*(2), 205–217. https://doi.org/10.51717/simkom.v9i2.478

Rui, L., & Xiuli, G. (2022). Basic Research on Construction of Multimodal Parallel Corpus of Tourism Translation in New Media Era. *Academic Journal of Humanities & Social Sciences*, *5*(15), 139–144. https://doi.org/10.25236/ajhss.2022.051519

Shen, N. (2022). English-Chinese Corpus Collection and Translation Wisdom Algorithm Implementation Based on Ajax+JQuery. *International Journal of Science and Engineering Applications*, *11*(12), 300–302. https://doi.org/10.7753/ijsea1112.1015

Spatioti, A. G., Kazanidis, I., & Pange, J. (2022). A Comparative Study of the ADDIE Instructional Design Model in Distance Education. *Information 2022, 13*, 1–22.

Sudibyo, B. (2008). *Tesaurus Bahasa Indonesia Pusat Bahasa*. Departemen Pendidikan Nasional.

Team. (2025). Si Palui. *Banjarmasin Post*.

Winda, N., & Muhammad, A. (2023). Pengembangan Parsing PCPATR sebagai Preservasi Bahasa dan Sastra Banjar. In *Jurnal Onoma: Pendidikan, Bahasa dan Sastra* (Vol. 9, Issue 2). Pendidikan. https://e-journal.my.id/onoma