Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666

Volume 8 Issue 33 (December 2023) PP. 178-198
DOI: 10.35631/JISTM.833014

Journal of Information Systems and Technology Management (JISTM)
eISSN: 0128-1666

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT (JISTM)
www.jistm.com

A MORPHOLOGICAL ANALYSIS OF MALAY TRANSLATED QUR'AN CORPUS

Nor Diana Ahmad[1], Nurhidayah Bahar[2*], Fadilla 'Atyka Nor Rashid[3], Ahmad Zambri Shahuddin[4]

[1] School of Computing Science, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
Email: diana12@uitm.edu.my
[2] Centre for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
Email: nbahar@ukm.edu.my
[3] Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
Email: fadilla@ukm.edu.my
[4] School of Computing Science, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
Email: azambri@uitm.edu.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 02.10.2023
Revised date: 30.10.2023
Accepted date: 21.11.2023
Published date: 10.12.2023

To cite this document:

Ahmad, N. D., Bahar, N., Rashid, F. 'A. N., & Shahuddin, A. Z. (2023). A Morphological Analysis Of Malay Translated Qur'an Corpus. *Journal of Information System and Technology Management, 8* (33), 178-198.

DOI: 10.35631/JISTM.833014.

Abstract:

Research in Natural Language Processing for English has shown successful results for more than a decade. However, it is difficult to adapt those techniques to the Malay language, because its complex morphology and orthographic forms are very different from English. Moreover, limited resources and tools for computational linguistic analysis are available for Malay. This study aims to demonstrate the use of Malay morphological analysis on the Malay translated Qur'an, for research and teaching purposes. To present the use of Malay morphological analysis, we propose a morphological analysis algorithm for the Malay-translated Qur'an that allows it to be created, discovered, and queried. This involved the creation of the first Malay-translated Qur'an corpus of 149,654 words with root word annotation and a new morphological analysis algorithm for Malay-translated Qur'an. Besides, a new list of Malay Stop words, a new rule-based stemming algorithm, and a new root word annotation were developed. This new Malay language resource will benefit other researchers, especially in religious morphological analysis and semantic modelling in research related to Malay documents.

Keywords:

Morphological Analysis, Natural Language Processing, Malay Translated Qur'an, Computational Linguistic

Introduction

The Malay language is part of the Austronesian language family, and it is widely used in the Southeast Asia region that includes Malaysia, Singapore, Indonesia, Brunei, and Thailand by over 306 million people based on the March 2020 Statistic on the Internet World Stats page. Malay is the national language of Malaysia, and it is used as the medium of communication in government schools and as an intra-ethnic communication tool (Omar, 2017). The Malay language is also known as "Bahasa Melayu" in the Malaysian context. Although traditional linguistics as such theories, models, and methodologies are well developed for the Malay language, limited resources, and tools are available or accessible for computational linguistic analysis (Ahmad et al., 2016).

A corpus is a methodically assembled compilation of electronic textual content designed to encapsulate language features pertinent to computational linguistic research to the greatest extent possible. The creation of most corpora is motivated by scientific imperatives essential for research endeavors. For example, the construction of a contemporary Arabic corpus, as undertaken by (Dukes et al., 2010), involves selecting texts that faithfully mirror the intricacies of the language. Consequently, the development of a corpus becomes imperative in order to accurately portray the linguistic resources essential for supporting advancements in research and technology within the language.

There has been a significant development in creating a corpus in some European and Asian countries in their native languages. However, there has yet to be any recognizable attempt to create a Malay corpus. The Dewan Bahasa dan Pustaka (DBP) Corpus is the only Malay corpus comprising 114 million words taken from various sources, from modern to classical Malay texts. Regrettably, this corpus is not publicly available and has limitations in studying modern Malay. For the Qur'an with Malay language translation, the words used are richer than other Malay document texts such as newspapers, reports, or books. This is because the Qur'an was originally written in Classical Arabic, and the translation becomes more challenging due to the unique style of writing and several different orthographic forms. As for the Malay-translated Qur'an corpus, the words used in the translation are taken from both modern and classical Malay. However, 286 out of 149,654 words used in this translation are derived from Arabic. Because of the multilanguage used in this document, there is a need to extract, compile, categorize, and annotate this corpus and reuse the word for other morphological analysis.

Thus, this study aims to demonstrate the use of Malay morphological analysis on the Malay-translated Qur'an for research and teaching purposes. To present Malay morphological analysis, we need to describe a morphological analysis algorithm for the Malay-translated Qur'an that allows it to be created, discovered, and queried. This involved the creation of the Malay-translated Qur'an Corpus and a new morphological analysis algorithm for the translated Qur'an. This includes a new list of Malay Stop words, a new rule-based stemming algorithm, and a new root word annotation.

Literature Review

Natural Language Processing (NLP)

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human languages. Natural language processing systems take strings of words (sentences) as their input and produce structured representations capturing the meaning of those strings as their output (Eggebraaten et al., 2014). There are two problems in processing natural language. These problems are the ambiguity level in natural languages and the complexity of semantic information contained even in simple sentences. The ambiguity of the query and returned results can be one of the major constraints. This ambiguity usually happens in understanding the actual meaning of the user's query and giving the right results. Dealing with the natural ambiguity of languages, such as polysemy and synonymy, the IR community moves from words that are expressions of natural language to concepts expressed in an ambiguous format in the form of formal language such as ontology, which is generally described as word sense disambiguation. With the emergence in the 1970s of models of ranked retrieval that process unstructured queries, automatic query systems became a fact. The central philosophy of automatic query systems is that indexing and query formulation should result in a representation closer to the text's actual meaning, ignoring as many of the irregularities of the natural language as possible. A typical indexing and query formulation approach selects the query terms as follows. First, a tokenization process occurs, then the stop words are removed, and finally, the remaining words are stemmed. Additionally, natural language processing modules might provide the identification of phrases or the splitting of compounds.

There has been a growing interest in Malay language processing and translation, but most of the researchers focused on a different set of problems, such as information retrieval documents (Bakar et al., 2003; Othman & Wahid, 2011; Syazhween et al., 2018; Yahya et al., 2013), build POS Tagger (Alfred et al., 2013; Hamzah, 2014; Mohamed et al., 2011; Mohd Don, 2010; Noor et al., 2018; Othman & Wahid, 2011; Xian et al., 2016), Malay Corpus construction and analysis [9, 22, 16, 32, 7](Bakar, 2020; Baldwin & Awab, 2006; Griffiths, 2018; Lee & Min, 2012; Rodzman et al., 2018) and ontology (Taa, 2012; Taa et al., 2018; Yahya et al., 2013). However, lexical coverages for corpus construction are usually obtained from blogs, newspapers, and other online resources (Taa et al., 2018). Limited research has been done on text processing in a document like the Qur'an. SEAlang Library Malay Text Corpus consists of Malay texts retrieved from various Internet Sources (Baldwin & Awab, 2006). However, it only focuses on collocation. Although studies have been conducted relevant to this matter, limited resources and tools for computational linguistic analysis are available for the Malay language. To the best of our knowledge, no corpus has been developed specifically on Malay-translated Qur'an or other Malay-translated religious text up until the present time.

Morphology in the Malay Language

Natural Morphology is the field of language structure, form, and classification of words. The word structure is the arrangement of the speech or symbolic form (written) built up from a smaller meaning-bearing unit called morphemes to become a meaningful language unit. Morphologically, Malay is a language that belongs to the agglutinative language family. It does a lot of affixation, reduplication, and composition (compounding) as well as other rarely used processes (such as deletion or producing acronyms) in word formation (Sharum et al., 2010).

The meaning of words can be changed by adding inflectional morphemes such as prefixes, suffixes, and circumfixes to the root words.

Affixation

Affixation is the process whereby a base word may be extended or added by one or more affixes. Affixation is the most common process of the three morphological processes. Affixes can be classified as prefixes, suffixes, infixes, and circumfixes. According to (Hassan, 2002), affixes are used to add a word's meaning. Table 5.4 shows the affixes, type of affixes, and category of words for every affix based on the rules of Alkhawarizmi word labelling.

Table 1: Affixes Labeling Based on Al Khwarizmi's Rule

Affixes	Type of Affixes	Category of words
pe	prefix	Noun
pen	prefix	Noun
pem	prefix	Noun
peng	prefix	Noun
penge	prefix	Noun
pe	prefix	Noun
per	prefix	Noun
ke	prefix	Noun
juru	prefix	Noun
be	prefix	Verb
bel	prefix	Verb
ber	prefix	Verb
Me	prefix	Verb
Men	prefix	Verb
Mem	prefix	Verb
meng	prefix	Verb
Menge	prefix	Verb
memper	prefix	Verb
Di	prefix	Verb
diper	prefix	Verb
ter	prefix	Verb
i	suffix	Verb
kan	suffix	Verb
diper+i	infix	Verb
me+kan	infix	Verb
me+i	infix	Verb
mem+kan	infix	Verb
mem+i	infix	Verb
meng+kan	infix	Verb
meng+i	infix	Verb
menge+kan	infix	Verb
menge+i	infix	Verb
memper+kan	infix	Verb
memper+i	infix	Verb

Studies in the affixes for the Malay language are relatively left behind in comparison to other languages such as English and European languages. The usage of affixes in English and other European languages is less complex than the Malay language as it has been found that the stemmers are only concerned with the removal of suffixes. However, in Malay morphology, a stemmed word is produced by removing affixes in the text, document, or query. Affix is the verbal element that attaches to the word whether at the beginning of the word (prefix) and the end of the word (suffix). Besides, more than one affix may also be attached to a word at the same time. The word also can contain both affixes and this is known as prefix-suffix pair, for example, as seen in the word ‘pemakanan’ (nutrition). The root word for this word is ‘makan’ (eat), and the prefix ‘pe’ is added at the beginning and the suffix ‘an’ at the end of the word to complete the word ‘pemakanan’.

On the other hand, English and Malay languages differ regarding their root words, which are based on their respective morphological structures (Abdullah, Ahmad, Mahmod, & Sembok, 2009). For instance, the English words ‘related’, ‘relates’, and ‘relation’, is derived from the root word ’relate’, and stemmer can work as suffix removal for the English language. Yet, the Malay language has a different stemming process than English due to the complexity of its morphological rules. For example, the Malay words ‘pengajaran’, ‘pembelajaran’, and ‘pelajar’ are derived from the root word ’ajar’, and it is insufficient to use suffix removal to decide on the perfect root word (R. Khan et al., 2019).

Reduplication

Reduplication is a word-formation process expressing meaning by repeating all or part of a word. This reduplication is widely used in many Malay texts. Reduplication is hardly found in other languages. There are 3 basic categories of Malay reduplication: full, partial, rhyming, and chiming reduplication. Another free-form reduplication is the reduplication group, where their formation is not yet clearly understood, thus undefined. Full reduplication involves a base word, complex word, or compound word. For example, ‘perempuan-perempuan’ (girls) come from base word ‘perempuan’ (girl), ‘pertubuhan-pertubuhan’ (organizations) come from the complex word per- + ‘tubuh’ (body) + -an (‘pertubuhan’ or organization), and ‘kakitangan-kakitangan’ (staff) come from compound word ‘kaki’ (leg) and ‘tangan’ (hand).

Partial reduplication is a process that reduplicates one part of the base or root word. It can be divided into first-syllabic reduplication, root reduplication, and compound reduplication. First, syllabic reduplication is a reduplication process that derives a noun from another noun, by reduplicating the first syllabic and converting the vowel of the copy to a letter ‘e’ (called ‘e-pepet’) such as ‘pepatung’ (dragonfly) from ‘pa’ + ‘patung’ (statue). Root reduplication is reduplicating the root word of the derived base word. The reduplication can be positioned in front of or behind the base word. This position usually reflects the meaning of the derived word. The root positioned at the front generates a meaning of a reciprocal act called ‘menyaling’ e.g, ‘pukul-memukul’ (hitting each other), while a reduplication root positioned at the back generates a meaning of multiple or repeated acts e.g, ‘berlari-lari’ (running). The difference between root word duplication and reciprocal is just the meaning where it describes repeated and opposing acts. Both are rooted in word duplication (Yunus, Zainuddin, & Abdullah, 2010). Compound reduplication is a partial reduplication when it follows the Distributed Morphology (DM) rules, where the main component of the compound word precedes the other component e.g, ‘alat tulis’ (stationery). Rhythmic reduplication involves repeating certain forms of the root word, such as consonants, syllables, or vowels, which creates symphonic sounds in the pronunciation. For example, ‘kayu-kayan’ (woods), ‘batu-batuan’ (stones) and ‘gunung-ganang’ (mountains). Free-form reduplication is a reduplication that does not belong to any of the categories above. For instance, ‘sahabat-handai’ (friends), ‘nenek-moyang’ (ancestors), and ‘ipar-duai’ (brothers and sisters-in-law).

Tokenization

This study also undertakes tokenization. Tokenization is the task of chopping text into pieces, also called tokens. In the early nineties, (Webster & Kit, 1992) published a paper on the significance and complexity of tokenization at the beginning of natural language processing. The authors provide their arguments from the perspective of lexicography and pragmatism. (Heinzerling & Strube, 2018) presented a study on tokenization-free pre-trained Subword Embeddings in 275 Languages, and in the following year, (Mullen et al., 2018) reported fast, consistent tokenization of natural language text. It can be seen that a tokenization process plays a role in improving the accuracy of part-of-speech (POS) tagging. Since the Malay-translated Qur’an uses many stop words in their translation text, eliminating them is also a necessary process.

Composition or Compounding

Compounding is linking two or more basic words into single words with a specific meaning. Compound words may be hyphenated, written open as separate, or solid. Most Malay compound words are constructed by nouns, and other nouns, verbs, and adjectives modified it. For instance, ‘adat-istiadat’ (customs and tradition), which linked a single word, ‘adat’ (customs), with another single word, ‘istiadat’ (tradition). This study will not focus on compound words. We concentrate on reduplication words as this word appeared a lot in the Malay translation of the Qur’an text.

A significant problem in the Malay morphological processing of Malay documents is the analysis part. This often happens when creating information retrieval applications (stemming) and corpus tagging (glossing). The morphological analysis process analyzes the lexical form of a word from its root form. For instance, to identify the root where the word originates in stemming or conflation and to identify the underlying word’s structure and features for word glossing in corpus tagging. In creating Malay Stemmers, not only do we need to remove the affixes, but understanding the variations of different aspects of the four affixes shown above is crucial. Although many researchers have found solutions, the under-stemming and over-stemming issues remain unresolved (Sharum et al., 2010). This also causes morphological analysis for the Malay language to be still far behind, and there are no accessible resources for morphological analysis results.

Data Collection and Analysis

Data collection is an essential part of corpus development. We were looking for a digital version of the Malay-translated Qur’an for this study. There are few digital versions of the Malay-translated Qur’an available online. However, no digital version of the Qur’an meets our criteria. Most Malay digital versions of the Qur’an translation combine Malay and Indonesian. Therefore, we decided to use the Al-Qur’an Amazing book published by Karya Bestari (Nursalim et al., 2016) as our dataset. The purpose of using this Malay Translated Qur’an is that the Malay words used in the translation have been reviewed and verified by JAKIM, Department of Islamic Development, Malaysia. Besides, it contains modern and Classical Malay, which are important in the corpus construction. The Qur’an is divided into 114 chapters

(Suras) of varying sizes, and each chapter is divided into verses (Ayahs). There are 6,234 verses in the Qur'an. The extraction process must be done since the data is from the book. In this study, the extraction process is done manually by extracting every chapter and verse. Because the extraction process is done manually, the error probability is high. Thus, data analysis becomes an ongoing iterative process where data is continuously collected and analyzed almost simultaneously.

Data Analysis

Data analysis is intended to validate the data gathered. Data analysis ensures that all chapters, verses, and words are correct. Two levels of validation have been done on the data for this study. The first validation is to analyze and compare the structure and content of Malay Qur'an translation with English and Arabic translations. This is to ensure that the structure and words used are synchronized. The second validation is by two experts in the field of Quranic studies. Both experts will ensure that the chapters, verses, and words used in the translation are correct. The Qur'an analyzed data is then stored in the Oracle 11g database. Table 2 shows the statistics of data extracted from the Malay-translated Qur'an book.

Table 2: Statistic of Extracted Data

Data	Statistics
Total number of chapters	114
Total number of verses	6,236
Total number of words	149,654

Methodology

One of the essential parts of morphological analysis is pre-processing. In this study, we analyze pre-processing to a Malay-translated Qur'an. This method prepares the Malay-translated Qur'an data and its structure for the data annotation process. Stop word removal, tokenization, reduplication, and stemming are the techniques used in this pre-processing phase. The importance of pre-processing in this study is:

a. Cleaning: This process is to remove unwanted parts of text such as punctuation marks, stop words, capital letters, and other characters that appear in the text.
b. Normalization: This process is important for the Information Retrieval process. It will retrieve the base form of the word by reducing the dimensionality of the size of the index words, usually through Stemming, tokenization, and other forms of standardization.
c. Analysis: This analysis process usually consists of statistical and visualization of data.

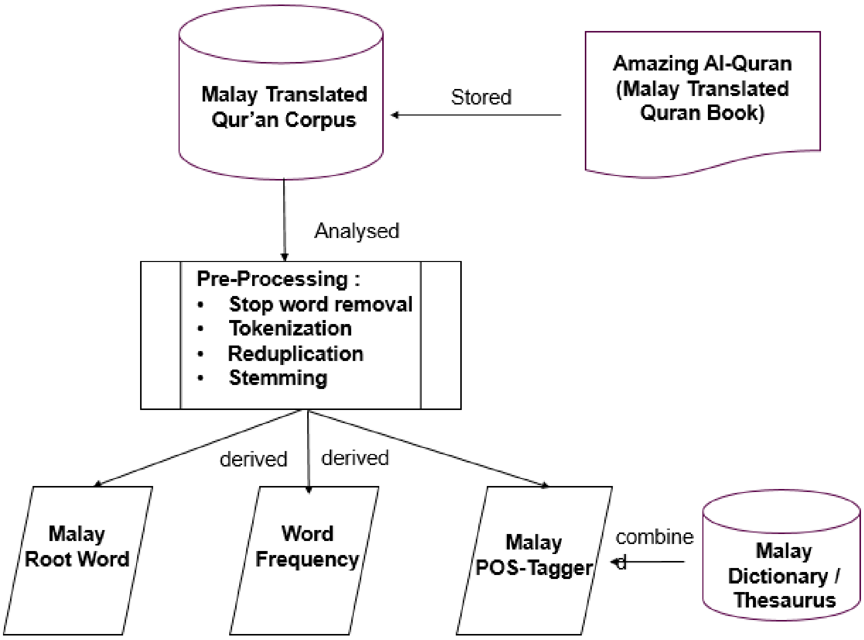
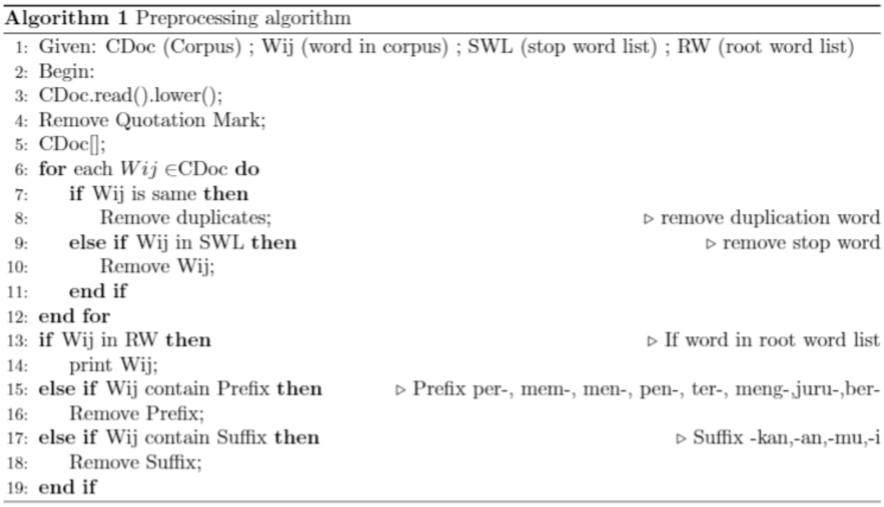


Figure 1: The Architecture Morphological Analysis Of The Malay-Translated Qur'an Corpus Development

The architecture depicted in Figure 1 illustrates the application of NLP techniques in processing Malay-translated Qur'an texts, specifically focusing on the morphological analysis of the Malay-translated Qur'an Corpus. This corpus incorporates the Malay translated Qur'an book. The initial phase of analysis involves the implementation of pre-processing techniques to optimize the efficiency and accuracy of subsequent investigations. Key pre-processing steps include removing stop words, tokenization for breaking down the text into individual words, addressing reduplication to handle repetitive patterns, and stemming to reduce words to their root forms. Anticipated deliverables post pre-processing include the identification of Malay root words, the analysis of word frequency, and the application of a Malay POS-tagger. These outcomes will culminate in the development of a Malay dictionary or thesaurus.

The text pre-processing is implemented using Python 3.4, as depicted in Figure 2, which illustrates the algorithm for pre-processing text.



```
Algorithm 1 Preprocessing algorithm
Given: CDoc (Corpus) ; Wij (word in corpus) ; SWL (stop word list) ; RW (root word list)
Begin:
CDoc.read().lower();
Remove Quotation Mark;
CDoc[];
for each $Wij$ ∈CDoc do
    if Wij is same then
        Remove duplicates;                              ▷ remove duplication word
    else if Wij in SWL then                             ▷ remove stop word
        Remove Wij;
    end if
end for
if Wij in RW then                                       ▷ If word in root word list
    print Wij;
else if Wij contain Prefix then     ▷ Prefix per-, mem-, men-, pen-, ter-, meng-,juru-,ber-
    Remove Prefix;
else if Wij contain Suffix then                         ▷ Suffix -kan,-an,-mu,-i
    Remove Suffix;
end if
```

Figure 2: Pre-processing Algorithm

The architecture portrays the natural language processing (NLP) techniques used in processing the Malay-translated Qur’an text. This study employs stop words removal, tokenization, and reduplication at the pre-processing stage. Subsequently, word frequency and part-of-speech (POS) taggers were derived from the pre-processing stage. Additionally, this study also employs concordance to examine relationships between words. Such methods and techniques are most appropriate for processing text in the chosen document.

Apart from these text processing methods, this study also uses Zipf’s Law. Zipf’s Law is a statistical distribution in particular data sets. In the case of our study, words in the digital Qur’an have been translated into the Malay language. We identified the frequencies of certain words and ranked them. A growing body of literature uses Zipf’s Law to examine the frequency of words in natural language (Bentz & Ferrer-i-Cancho, 2016; N. et al., 2019; Yatsko, 2015). Authors argue that Zipf’s Law can significantly reduce text dimensionality and improve text classification automation’s speed (Yatsko, 2015).

Results and Discussion

This section presents the analysis and the result based on using Natural Language processing techniques in the pre-processing stage. This finding is essential in presenting the use of Malay-translated Qur’an corpus in different computational linguistic research. The Malay-translated Qur’an has been processed during the pre-processing phase to remove the noise. Stop words removal, tokenization, and reduplication are the techniques used in this phase.

Capitalization and Character removal

Firstly, the words in the text are formatted in lowercase. Then, the punctuation marks need to be removed. Many punctuation marks are used in Malay text, such as ’!’, ’?’, ’(’, ’)’, ’.’, ’“’, ’*’, ’”’, ’,’. Removing punctuation marks is not easy since punctuation marks that mark the end of a sentence are often ambiguous. To disambiguate punctuation marks often relies on regular expressions. Here, regular expression rules are applied to remove the punctuation marks.

Split the Reduplication Words

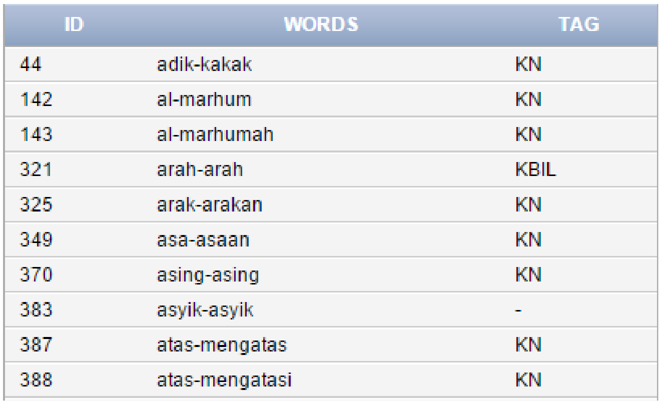
Secondly, splitting reduplication words appeared in this corpus. Reduplication is a word-formation process in which meaning is expressed by repeating all or parts of a word. Splitting the reduplication words is important to gather the root words. The reduplication word is widely used in many Malay texts. Reduplication is rare in other languages. Reduplication is usually found in Austronesian languages such as Malay. Reduplication indicates the simple plural or 'many'. In Malay, if the user knows there is more than one of an object but does not know or does not wish to specify how many, the whole form of the noun may be repeated twice to signal the plural. For example, if there is more than one cat or 'kucing', you will say a cat or 'kucing-kucing' in Malay. Generally, reduplication in Malay can be divided into full, such as 'perempuan-perempuan' or girls (from 'perempuan' or girl), or partial, such as 'lelaki' or boys (from 'laki-laki' or boys) or rhyming and chiming, such as the word 'kayu' or wood combines with 'kayan' or wood to form 'kayu-kayan' different sorts of wood.

In this Malay Translated Qur'an, we found 1,587 words in reduplication. According to Malay linguists (Mohd Don, 2010), only the first word is considered the root, and the second should be removed. However, the second word in this study will not be discarded as some of the second words are root words. Considering this scenario, the reduplication process is done in the semiautomatic process. First, we used Python code to split the reduplication words. After that, we checked each word manually and indicated whether the word was a root word. Figure 3 shows the list of reduplication words found in the Malay translated Qur'an text, while Figure 4 shows the results after splitting the reduplication words. Below is the example of reduplication that appeared in the Malay-translated Qur'an and the English-translated Qur'an. Based on the sample given in Table 3, it can be observed that many plural nouns in the Malay language use double words.

- *Malay Translated Qur'an: "Dan apabila mereka bertemu dengan* ***orang-orang*** *yang beriman, mereka berkata, "Kami telah beriman". Tetapi apabila mereka kembali kepada* ***syaitan-syaitan*** *(para pemimpin) mereka, mereka berkata,"Sesungguhnya kami bersama kamu, kami hanya* ***berolok-olok****."" (Al-Baqarah, 14)*
- *English Sahih International: "And when they meet those who believe, they say, "We believe"; but when they are alone with their evil ones, they say, "Indeed, we are with you; we were only mockers." (Al-Baqarah, 14)*

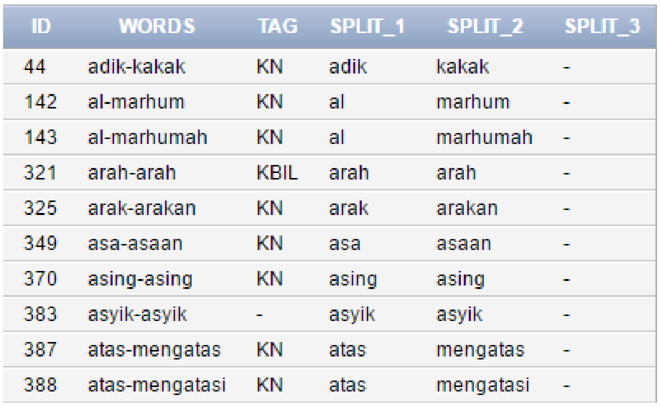
Table 3: Statistic of Extracted Data

Malay	English	Root word in Malay
Orang-orang	Those (human)	Orang
Syaitan-syaitan	Devils (evil ones)	Syaitan
Berolok-olok	Mockers	Olok



ID	WORDS	TAG
44	adik-kakak	KN
142	al-marhum	KN
143	al-marhumah	KN
321	arah-arah	KBIL
325	arak-arakan	KN
349	asa-asaan	KN
370	asing-asing	KN
383	asyik-asyik	-
387	atas-mengatas	KN
388	atas-mengatasi	KN

Figure 3: Example of Reduplication Words in Corpus

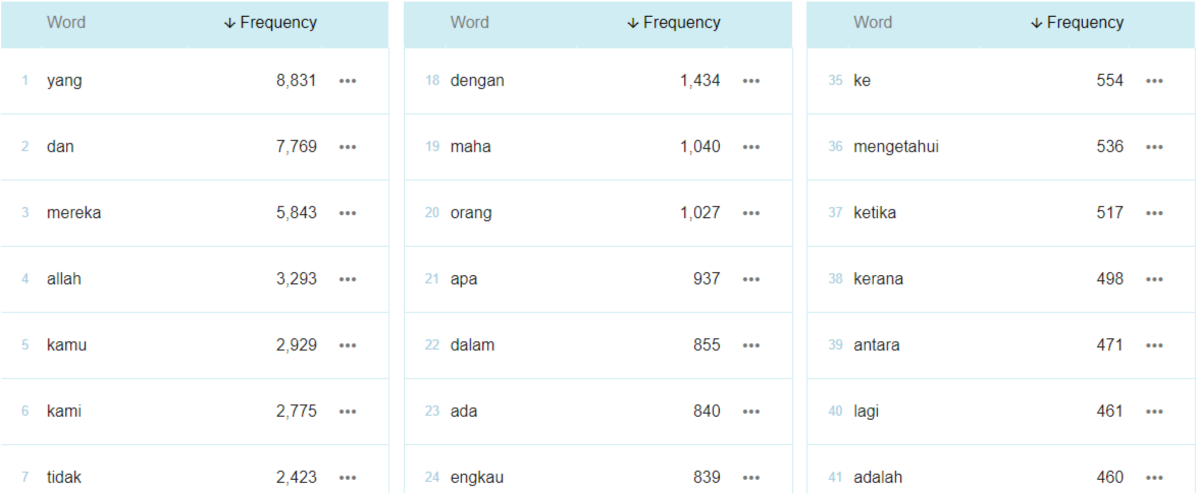


ID	WORDS	TAG	SPLIT_1	SPLIT_2	SPLIT_3
44	adik-kakak	KN	adik	kakak	-
142	al-marhum	KN	al	marhum	-
143	al-marhumah	KN	al	marhumah	-
321	arah-arah	KBIL	arah	arah	-
325	arak-arakan	KN	arak	arakan	-
349	asa-asaan	KN	asa	asaan	-
370	asing-asing	KN	asing	asing	-
383	asyik-asyik	-	asyik	asyik	-
387	atas-mengatas	KN	atas	mengatas	-
388	atas-mengatasi	KN	atas	mengatasi	-

Figure 4: Result After Splitting the Reduplication Words

Tokenization

The following process is tokenization. The purpose of tokenization is to split a text into meaningful units called tokens. This research requires tokenization to gather root words to facilitate information retrieval (Ahmad, 1995). One hundred forty-nine thousand six hundred fifty-four tokens have been extracted from the corpus—the extraction process used two software programs, Sketch Engine and Nvivo 10. Figure 5 shows an example of tokens derived from Malay translated Qur'an.



	Word	↓ Frequency	
1	yang	8,831	•••
2	dan	7,769	•••
3	mereka	5,843	•••
4	allah	3,293	•••
5	kamu	2,929	•••
6	kami	2,775	•••
7	tidak	2,423	•••

	Word	↓ Frequency	
18	dengan	1,434	•••
19	maha	1,040	•••
20	orang	1,027	•••
21	apa	937	•••
22	dalam	855	•••
23	ada	840	•••
24	engkau	839	•••

	Word	↓ Frequency	
35	ke	554	•••
36	mengetahui	536	•••
37	ketika	517	•••
38	kerana	498	•••
39	antara	471	•••
40	lagi	461	•••
41	adalah	460	•••

Figure 5: The Results from the Tokenization Process

Stop Word Removal

The elimination of stop words is performed. A stop word is a word that does not carry meaning in natural language and, therefore, can be ignored. Stop words are commonly eliminated from many text processing applications because these words can be distracting, non-informative, and have additional memory overhead. Removing stop words from the corpus also leads to its decreased size, which increases the efficiency of any NLP activity (Raulji & Saini, 2017). In Malay language, there are many stop words such as ‘adalah’, ‘yang’, ‘itu’, ‘selepas’ and ‘mereka’ (Al-Saffar Ahmed & Awang, 2018). These stop words and others are abundant in the Malay language. Thus, these stop words must be removed to omit the text's low-level information and ultimately focus on the critical information. Furthermore, removing the stop words could help reduce training time because of a smaller dataset with fewer tokens involved. In this study, we used 356 stop words from (Ahmad, 1995). However, after going through the stop word, we found several duplicate and irrelevant words. Thus, we have to eliminate those words. Therefore, the total number of stop words used in this research is 318 words.

Surprisingly, the most frequent words in this study after the tokenization process are stop words; therefore, half of the words appearing in a text do not need to be considered. This allows, for instance, a significant reduction in the space overhead of indexes for natural language text. Most of the Malay translated Qur’an text stop words are connection parts of a sentence rather than showing the subject, object, or intent. Table 4 shows the analysis result before and after the elimination of the stop words based on the seven (7) longest chapters in the Qur’an. As we can see from the results, the number of words containing stop words is more than those with subjects or objects.

Table 4: The Analysis Result of Before and After Elimination of Stop word

Chapter number / Documents	Total number of words	Total number of words (After removing the stop words)	Total number of words (contains stop words)
(2) Surah Al-Baqarah (The Opening)	94	4848	6446
(4) Surah An-Nisa (The Women)	7357	3173	4184
(3) Surah Ali' Imran (The Family of Imran)	6548	2770	3778
(7) Surah Al-A'raf (The Heights)	6538	2801	3737
(9) Surah At-Taubah (The Repentance)	4918	2071	2847
(10) Surah Yunus (Jonah)	3387	1394	1993
(12) Surah Yusuf (Joseph)	*3357*	1455	1902
TOTAL	**43,399**	**18,512**	**28,887**

Stemming

Stemming is a computational process of reducing a word from its derived form into its root term. Stemming helps improve retrieval performance by reducing variants of the same root word to a familiar concept. Furthermore, stemming has the secondary effect of reducing the size of the indexing structure because the number of distinct index terms is reduced. The stemming algorithm reduces words with the same root to a standard form by stripping each word of its derivational and inflectional suffixes (Syed Abdullah, 2012). Malay language affixes consist of four different types of verbal elements:

1. Prefix: attaches itself at the beginning of a word. Example: *'bersalah'*, which means guilty, starts with *'ber', a typical Malay prefix*. Other prefixes appeared in Malay, such as *'per'*, *'mem'*, *'men'*, *'pen'*, 'ter', 'meng', and 'juru'.
2. Suffix: attaches itself at the end of a word. Example: *'memaafkan'*, which means to forgive, ends with '*kan', a typical Malay suffix*. Another suffix is *'an'*, *'i'*, and *'mu'*.
3. Infix: usually located in the middle of the word. Example: *'gerigi' means toothed blade,* derived from the root word *'gigi'* (teeth).
4. Circumfix: Prefix-suffix pair where more than one affix is attached to a word simultaneously and usually positioned before and after the root word. Example: *'kerajaan'*, which means kingdom, is derived from the root word *'raja'* (king).

Below is an example of affixes in the Malay language.

- Root word: *'cipta'* (create)
- Prefix: *'tercipta'* (created), *'mencipta'* (create)
- Suffix: *'ciptaan'* (creation), *'ciptaannya'* (his/her creation)
- Circumfix: *'menciptakan'* (create), *'menciptakanku'* (created me), *'menciptakanmu'* (created you), *'menciptakannya'* (created it), *'menciptanya'* (create it)

Volume 8 Issue 33 (December 2023) PP. 178-198
DOI: 10.35631/JISTM.833014

The proposed arrangement of the rules applied in the stemming process can be described as follows in Figure 6:

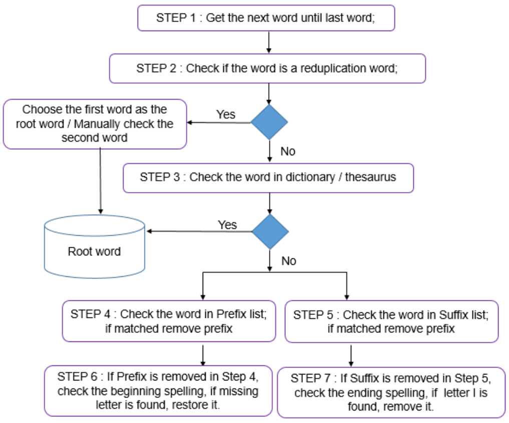

STEP 1 : Get the next word until last word;
STEP 2 : Check if the word is a reduplication word;
Yes
Choose the first word as the root word / Manually check the second word
No
STEP 3 : Check the word in dictionary / thesaurus
Yes
Root word
No
STEP 4 : Check the word in Prefix list; if matched remove prefix
STEP 5 : Check the word in Suffix list; if matched remove prefix
STEP 6 : If Prefix is removed in Step 4, check the beginning spelling, if missing letter is found, restore it.
STEP 7 : If Suffix is removed in Step 5, check the ending spelling, if letter I is found, remove it.

Figure 6: Flow Diagram of Stemming Rules

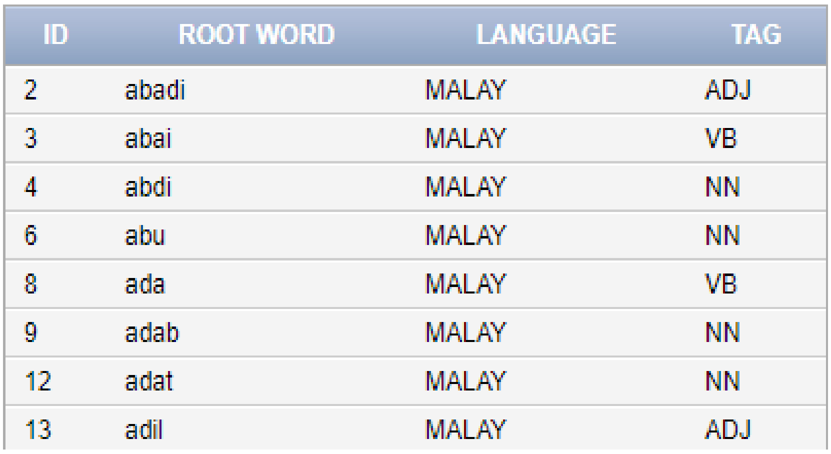
Based on this algorithm, we managed to stem 2,187 root words. Table 4 shows the statistics of words before and after the stop word removal and stemming processes. As we can see, the number of words after the stemming process is relatively low compared to before. This is due to many affixes in a root word such as 'mem-', 'ber-', 'meng-', '-nya', '-kan', and '-I' in Malay translated Qur'an. Table 5 shows the number of occurrences of affixes used in the Malay-translated Qur'an. According to the table, a circumfix is a group of words widely used in Malay translated Qur'an.

Table 5: Statistic of The Word in Malay Translated Qur'an Corpus

Words in Corpus Total	Total
Total number of words with stop word	149,654
Total number of words without stop word	63,191
Total number of words after stemming process and without stop word	2,187

Table 6: Affixes Used in Corpus.

Affixes used in Corpus	Number of Occurrence
Prefix	10,415
Suffix	7,672
Infix	53
Circumfix	11,800
TOTAL	**63,191**



ID	ROOT WORD	LANGUAGE	TAG
2	abadi	MALAY	ADJ
3	abai	MALAY	VB
4	abdi	MALAY	NN
6	abu	MALAY	NN
8	ada	MALAY	VB
9	adab	MALAY	NN
12	adat	MALAY	NN
13	adil	MALAY	ADJ

Figure 7: Stemming Result Stored in Oracle Database.

Figure 7 shows the stemming results stored in the Oracle Database. From 63,191 words containing affixes, only 2,817 words are root words. This shows that the Malay language uses many affixes in constructing a sentence. Each word of the affix describes a different meaning. The study by (Ahmad, 1995) and (Abdullah et al., 2009) showed that retrieval accuracy can be enhanced if the stop word is removed and stemmed from a word. All the stem words can be used as training resources for other research related to the Malay language, such as semantic annotation, POS-Tagging, Information Retrieval, and Semantic Modelling.

Word Frequency

A word frequency distribution is a way to display the most frequently occurring words or concepts in the document. To analyze the word frequency, Zipf's law has been employed. Zipf's law states that the most frequent word in a document will occur about twice as the second most frequent word, three times as often as the third most frequent word (Tullo & Hurford, 2003). For example, in the Malay translated Qur'an, the most frequently occurring word, "yang" appeared 12,615 times, accounting for nearly 6.24% of all the words (12,615 out of 174,457 occurrences). Aligned with Zipf's Law, the second-place word 'dan' accounts for slightly over 4.66% of words (9422 occurrences), followed by 'mereka' (6796 occurrences). All these words are stop words. Figure 8 shows the distribution of frequency rank in Malay translated Corpus based on Zipf's Law.

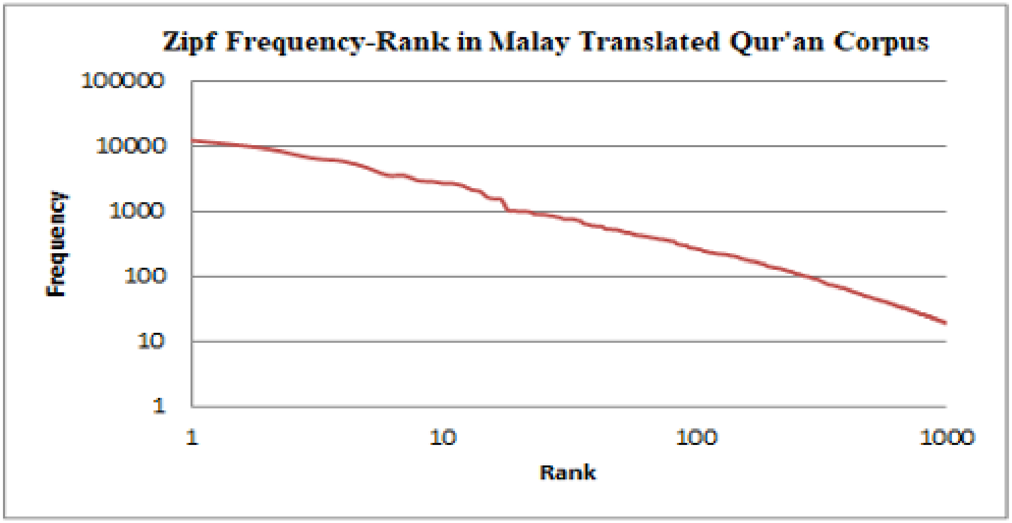


Figure 8: Zipf Frequency-Rank in Malay Translated Qur’an Corpus.

Part-of-speech (POS) Tagging

POS-tagging is a process of tagging a text or sentence into equivalent parts of speech tags based on the word definition and relation. A simple rule-based POS tagger applies to identify the word definition and relation. As for Malay POS Tagger, POS tag dictionary and affixing have been applied. The POS tag dictionary is manually extracted from Malay Dewan & Pustaka Thesaurus (Taharin, 2005) and stored in a database. Another resource used in this extraction process is the Online Malay POS Tagger by a Ph.D. student from the University of Malaya, Malaysia (Xian et al., 2016). Both resources helped in annotating the Malay-translated Qur’an text. The POS tags used in this research include Kata Kerja (Verb), Adjektif (Adj), Kata Tugas (Function), and Kata Nama (Noun). The POS tags used in this research such as Kata Kerja (Verb), Adjektif (Adjective), Kata Tugas (Function) and Kata Nama (Noun), and others such as Kata Bantu (Auxiliary), Kata Negatif (Negative), Nombor Kardinal (Cardinal Number). There are about 2,187 words that have been tagged in this POS-Tagger. Table 7 and Figure 9 show the frequency and relative frequency tagging for each POS tag in the Malay-translated Qur’an corpus. Nouns and verbs are the most common tag names in Malay-translated Qur’an documents.

Table 7: The Frequency and Relative Frequency in Malay translated Qur'an.

Tag	Tag Name	Frequency in Corpus	Relative Frequency (%)
NN	Kata Nama (Noun)	1179	53.91%
VB	Kata Kerja (Verb)	512	23.41%
ADJ	Adjektif (Adjective)	411	18.79%
FT	Kata Tugas (Function word)	60	2.74%
OTR	Others	25	1.14%

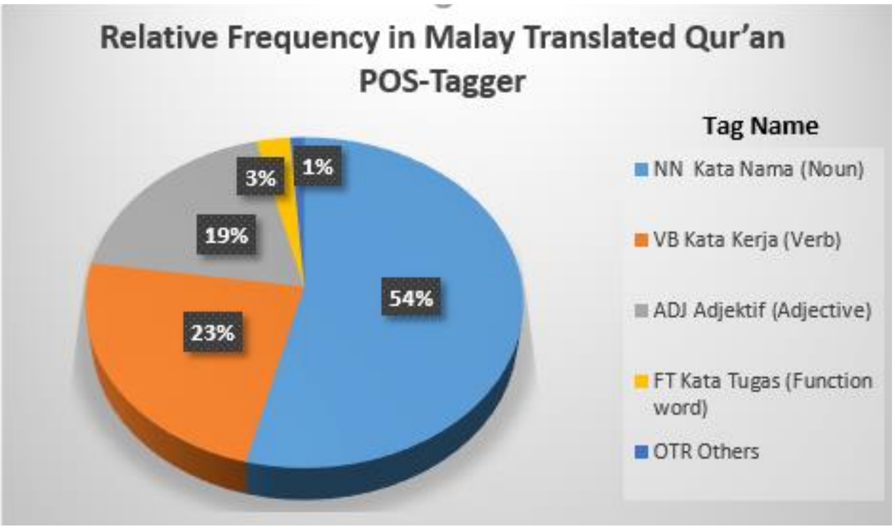
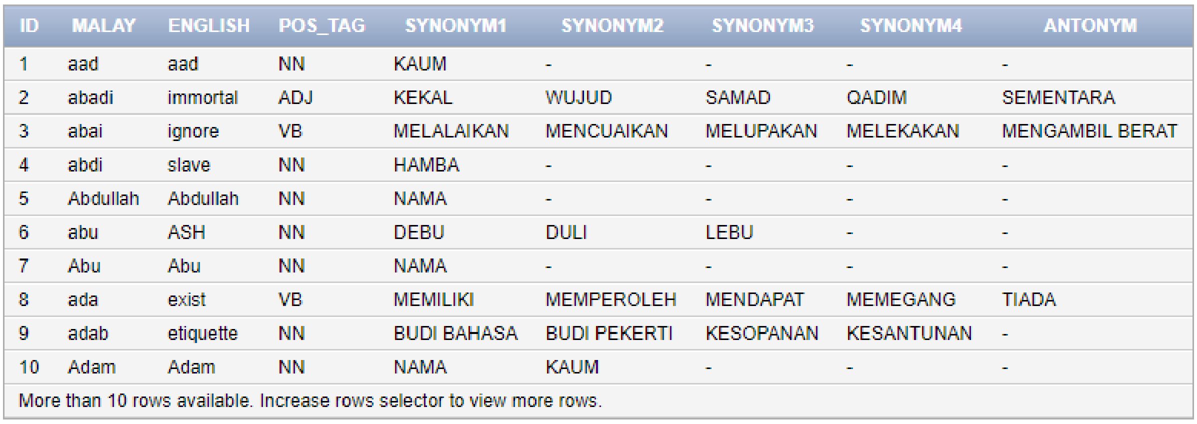


Figure 9: Pie Chart of The Relative Frequency of Malay Translated Qur'an POSTagger

Root Word Annotation

This phase is the core of the study, where the semantic annotation process is carried out for each word in the corpus. The 2,817 root words found in the corpus have been annotated. The annotation processes that we use for this study are synonyms and antonyms. These processes were made from scratch due to a need for more digital resources and references. The semantic annotation of synonyms is manually built using Malay Thesaurus by Dewan Bahasa dan Pustaka (DBP), Malay Dictionary, and WordNet. These semantic annotations can be used to create a concept for semantic relationship modeling.

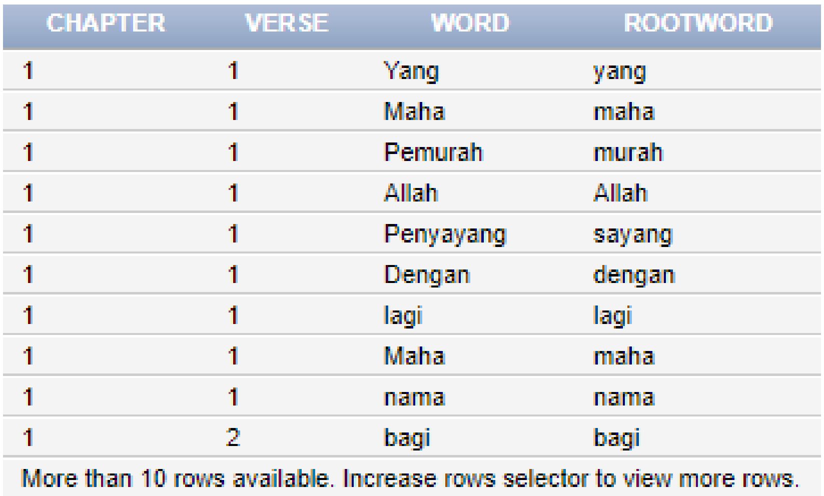


ID	MALAY	ENGLISH	POS_TAG	SYNONYM1	SYNONYM2	SYNONYM3	SYNONYM4	ANTONYM
1	aad	aad	NN	KAUM	-	-	-	-
2	abadi	immortal	ADJ	KEKAL	WUJUD	SAMAD	QADIM	SEMENTARA
3	abai	ignore	VB	MELALAIKAN	MENCUAIKAN	MELUPAKAN	MELEKAKAN	MENGAMBIL BERAT
4	abdi	slave	NN	HAMBA	-	-	-	-
5	Abdullah	Abdullah	NN	NAMA	-	-	-	-
6	abu	ASH	NN	DEBU	DULI	LEBU	-	-
7	Abu	Abu	NN	NAMA	-	-	-	-
8	ada	exist	VB	MEMILIKI	MEMPEROLEH	MENDAPAT	MEMEGANG	TIADA
9	adab	etiquette	NN	BUDI BAHASA	BUDI PEKERTI	KESOPANAN	KESANTUNAN	-
10	Adam	Adam	NN	NAMA	KAUM	-	-	-

More than 10 rows available. Increase rows selector to view more rows.

Figure 10: Example of Semantic Annotation Data Stored in Oracle Database.

Root word annotation is one of the contributions in this study. In linguistics, a root word holds the most basic meaning of any word. A root word has no suffix or prefix; it is the heart of the word. The root word annotation uses a list of root words derived from the affixing process. In natural language processing, it is essential to find the real root of a word for information retrieval and document categorization. In this study, we annotated 149,654 words with their root words, synonyms, and antonyms. Each root word is annotated according to the word's position in each chapter and verse in Malay translated Qur'an. Figure 11 shows the Malay translated Qur'an Corpus with root word annotation.



CHAPTER	VERSE	WORD	ROOTWORD
1	1	Yang	yang
1	1	Maha	maha
1	1	Pemurah	murah
1	1	Allah	Allah
1	1	Penyayang	sayang
1	1	Dengan	dengan
1	1	lagi	lagi
1	1	Maha	maha
1	1	nama	nama
1	2	bagi	bagi

More than 10 rows available. Increase rows selector to view more rows.

Figure 11: Malay Translated Qur'an Corpus with Root Word Annotation.

Conclusion

This study successfully establishes an annotated corpus for the Malay translated Qur'an, providing a groundwork for advancements in question-answer search and ontology development within the Qur'anic context. The utilization of morphological analysis techniques has yielded valuable insights, contributing to a nuanced understanding of pre-processing methods specific to Malay-translated Qur'anic texts. Theoretical contributions encompass various facets, including morphological processing methods, nuances in Malay language translation, and insights into Qur'an translation studies. Acknowledging certain limitations in the study, it is important to note that while numerous text processing methods exist, the focus has been on six specific ones: concordance, stop word removal, tokenization, reduplication, word frequency, and part-of-speech (POS) tagging. Despite the selectivity, these methods have proven most relevant for the chosen corpus, effectively achieving the study's objectives. Future research could explore additional text processing methods, such as collocation, term frequency-inverse document frequency, topic analysis, intent classification, and language classification. Expanding the scope to include other Qur'an translations in various languages would enhance the comprehensiveness of findings. Researchers are encouraged to delve deeper into these avenues, fostering the development of theories, models, and frameworks in this evolving field.

Acknowledgement

The authors express gratitude for the financial support provided by the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, which significantly contributed to the publication of this paper.

References

Abdullah, M., Ahmad, F., Mahmod, R., & Sembok, T. (2009). *Rules Frequency Order Stemmer for Malay Language*.

Ahmad, F. (1995). *A Malay Language Document Retrieval System: An Experimental Approach and Analysis*. (PhD thesis). Universiti Kebangsaan Malaysia (UKM), Malaysia.

Ahmad, N., Bennett, B., & Atwell, E. (2016). *Semantic-based Ontology for Malay Qur'an Reader*. Retrieved from https://api.semanticscholar.org/CorpusID:59378628

Alfred Rayner and Mujat, A. and O. J. H. (2013). A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles. In N. T. and H. H. Selamat Ali and Nguyen (Ed.), *Intelligent Information and Database Systems* (pp. 50–59). Berlin, Heidelberg: Springer Berlin Heidelberg.

Al-Saffar Ahmed AND Awang, S. A. N. D. T. H. A. N. D. O. N. A. N. D. A.-S. W. A. N. D. A. M. (2018). Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. *PLOS ONE*, *13*(4), 1–18. https://doi.org/10.1371/journal.pone.0194852

Bakar, N. S. A. A. (2020). The Development of an Integrated Corpus for Malay Language. In Y. and H. H. and O. C. K. Alfred Rayner and Lim (Ed.), *Computational Science and Technology* (pp. 425–433). Singapore: Springer Singapore.

Bakar Zainab Abu and Rahman, N. A. (2003). Evaluating the Effectiveness of Thesaurus and Stemming Methods in Retrieving Malay Translated Al-Quran Documents. In H. B. and C. H. and U. S. R. and M. S.-H. Sembok Tengku Mohd Tengku and Zaman (Ed.), *Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access* (pp. 653–662). Berlin, Heidelberg: Springer Berlin Heidelberg.

Baldwin, T., & Awab, S. (2006). Open Source Corpus Analysis Tools for Malay. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/677_pdf.pdf

Bentz, C., & Ferrer-i-Cancho, R. (2016). *Zipf's law of abbreviation as a language universal*. https://doi.org/10.15496/publikation-10057

Dukes, K., Atwell, E., & Sharaf, A.-B. M. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/278_Paper.pdf

Eggebraaten, T. J., Stevens, R. J., & Will, E. W. (2014). *Natural language processing ('NLP-overview')*. US Patent 8,639,495, 1–19.

Griffiths, A. (2018). The Corpus of Inscriptions in the Old Malay Language. In D. Perret (Ed.), *Writing for Eternity: A Survey of Epigraphy in Southeast Asia* (pp. 275–283). École française d'Extrême-Orient. Retrieved from https://hal.science/hal-01920769

Hamzah, M. P. Bin. (2014). *PART OF SPEECH TAGGER FOR MALAY LANGUAGE BASED ON WORDS MORPHOLOGY*. Retrieved from https://api.semanticscholar.org/CorpusID:16500895

Hassan, A. (2002). *Tatabahasa bahasa Melayu : morfologi dan sintaksis untuk guru dan pelajar*. Retrieved from https://api.semanticscholar.org/CorpusID:60244931

Heinzerling, B., & Strube, M. (2018). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, … T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1473

Khan, N., Bakht, M. P., Khan, M. J., Samad, A., & Sahar, G. (2019). Spotting Urdu Stop Words By Zipf's Statistical Approach. *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, 1–5. https://doi.org/10.1109/MACS48846.2019.9024817

Khan, R., Mohamad, F. S., Inam Ul Haq, M., Ahmad Zaidi Adruce, S., Adruce, Z., Anding, P., … Saleh, A. (2019). *Malay Language Stemmer*.

Lee, L. W., & Min, H.-T. (2012). *Developing an online Malay Language Word Corpus for primary schools*. Retrieved from https://api.semanticscholar.org/CorpusID:56566884

Mohamed, H., Omar, N., & Aziz, M. J. A. (2011). Statistical malay part-of-speech (POS) tagger using Hidden Markov approach. *2011 International Conference on Semantic Technology and Information Retrieval*, 231–236. https://doi.org/10.1109/STAIR.2011.5995794

Mohd Don, Z. (2010). Processing natural malay texts: A data-driven approach. *Trames Journal of the Humanities and Social Sciences*, *1464*, 90–103. https://doi.org/10.3176/tr.2010.1.06

Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software*, *3*, 655. https://doi.org/10.21105/joss.00655

Noor Ariffin, S. N. A., & Tiun, S. (2018). Part-of-Speech Tagger for Malay Social Media Texts. *GEMA Online Journal of Language Studies*, *18*, 124–142. https://doi.org/10.17576/gema-2018-1804-09

Nursalim, H., Muhtarom, A., Febriadi, S. R., Sanusi, F., Nurhakim, H., Fauzi, H., & Rahman, A. S. (2016). *Al-Quran Amazing.* (6th Edition). Shah Alam: Karya Bestari.

Omar, A. H. (2017). Languages and language situation of Southeast Asia. *Journal of Modern Languages*, *13*(1), 17–35. Retrieved from https://ejournal.um.edu.my/index.php/JML/article/view/3466

Othman, R., & Wahid, F. A. (2011). *Issues in Evaluating the Retrieval Performance of Multiscript Translations of Al-Quran*. Retrieved from https://api.semanticscholar.org/CorpusID:15945992

Raulji, J. K., & Saini, J. R. (2017). Generating Stopword List for Sanskrit Language. *2017 IEEE 7th International Advance Computing Conference (IACC)*, 799–802. https://doi.org/10.1109/IACC.2017.0164

Rodzman, S. B., Izuan Abdul Ronie, M. F., Ismail, N. K., Rahman, N. A., Ahmad, F., & Nor, Z. M. (2018). Analyzing Malay Stemmer Performance Towards Fuzzy Logic Ranking Function on Malay Text Corpus. *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 1–6. https://doi.org/10.1109/INFRKM.2018.8464767

Sharum, M. Y., Hamzah, Z. A. Z., Wahab, M. R. Abd., & Ismail, M. R. (2010). Formal Properties and Characteristics of Malay Rhythmic Reduplication. *Procedia - Social and Behavioral Sciences*, *8*, 750–756. https://doi.org/https://doi.org/10.1016/j.sbspro.2010.12.104

Syazhween Binti Zulkefli, N. S., Abdul Rahman, N. B., Puteh, M. B., & Abu Bakar, Z. B. (2018). Effectiveness of Latent Dirichlet Allocation Model for Semantic Information Retrieval on Malay Document. *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 1–5. https://doi.org/10.1109/INFRKM.2018.8464782

Syed Abdullah, F. (2012). SIMPLE RULES MALAY STEMMER. *The International Conference on Informatics and Applications (ICIA2012)*, 28–35.

Taa, A. (2012). *Ontology Development for ETL Process Design*. https://doi.org/10.4018/978-1-4666-1993-7.ch015

Taa, A., Abdullah Abed, Q., & Ahmad, M. (2018). Al-Quran ontology based on knowledge themes. *Journal of Fundamental and Applied Sciences*, *9*, 800. https://doi.org/10.4314/jfas.v9i5s.57

Taharin, M. (2005). *Tesaurus Bahasa Melayu Dewan.* . Kuala Lumpur, (Dewan Bahasa dan Pustaka).

Tan, T. P., Ranaivo-Malançon, B., Besacier, L., Yeong, Y.-L., Gan, K. H., & Tang, E. K. (2017). Evaluating LSTM Networks, HMM and WFST in Malay Part-of-Speech Tagging. *Journal of Telecommunication, Electronic and Computer Engineering*, *9*, 79–83. Retrieved from https://api.semanticscholar.org/CorpusID:67261520

Tullo, C., & Hurford, J. (2003). Modelling Zipfian distributions in language. *Proceedings of Language Evolution and Computation Workshop/Course at ESSLLI*, 62–75.

Webster, J. J., & Kit, C. (1992). Tokenization as the Initial Phase in NLP. *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*, 1106–1110. USA: Association for Computational Linguistics. https://doi.org/10.3115/992424.992434

Xian, B., Lubani, M., Liew, K., Bouzekri, K., Mahmud, R., & Lukose, D. (2016). Benchmarking Mi-POS: Malay Part-of-Speech Tagger. *International Journal of Knowledge Engineering*, *2*, 115–121. https://doi.org/10.18178/ijke.2016.2.3.064

Yahya, Z., Abdullah, M. T., Azman, A., & Kadir, R. A. (2013). Query Translation using Concepts Similarity Based on Quran Ontology for Cross-Language Information Retrieval. *Journal of Computer Science*, *9*(7), 889–897. https://doi.org/10.3844/jcssp.2013.889.897

Yatsko, V. A. (2015). Automatic text classification method based on Zipf's law. *Automatic Documentation and Mathematical Linguistics*, *49*(3), 83–88. https://doi.org/10.3103/S0005105515030048

Yunus, M. A., Zainuddin, R., & Abdullah, N. (2010). Semantic query with stemmer for Quran documents results. *2010 IEEE Conference on Open Systems (ICOS 2010)*, 40–44. https://doi.org/10.1109/ICOS.2010.5720061