*Universiti Malaysia Terengganu Journal of Undergraduate Research* *e-ISSN: 2637-1138*
*Volume 1 Number 3, Julai 2019: 87-95* 

# MALAY SINGLE SENTENCE PARSER

MUHAMMAD IQBAL NORDIN AND NOOR HAFHIZAH ABD RAHIM*

*School of Informatics and Applied Mathematics Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu*

*Corresponding author: *noorhafhizah@umt.edu.my*

**Abstract:** Parser is a process of classifying sentence structures of a language. Parser receives a sentence and breaks it up into correct phrases. The purpose of this research is to develop a Malay single sentence parser that can help primary school students to learn Malay language according to the correct phrases. This is because research in Malay sentence parsing has not gotten enough attention from researchers to the extent of building parser prototypes. This research used top-down parsing technique, and grammar chosen was context-free grammar (CFG) for Malay language. However, to parse a sentence with correct phrase was a difficult task due to lack of resources for obtaining Malay lexicon. Malay lexicon is a database that stores thousands of words with their correct phrases. Therefore, this research developed a Malay lexicon based on an article from *Dewan Masyarakat* magazine. In conclusion, this research can provide help to the primary school students to organize correct Malay single sentences.

Keywords: Malay single sentence parser, Malay language, Malay Lexicon

## Introduction

In Malaysia, research in the formulation of sentences describing the Malay texts still has not gotten enough attention from the researchers to the extent of building prototypes as it has with English (Yusnita and Zulikha, 2012). The word parser is a process of classifying the structure in the order of a sentence of a language (Mohanty and Balabantaray, 2003). The parsing process generates a useful parsing tree in grammar check applications. This application is the same as used in word processing systems (Hamden, 2012).

A single Malay sentence is a sentence that contains one subject and one predicate. A subject is the thing to describe which refers to people, things and places. A predicate is a group of words in a phrase that works to explain the subject. Single sentences consist of various types of sentences, whether statements, commands, verses and sentences.

According to Baidura and Jamilah (2011), a study of Malay language learning among speakers found that students committed 2,402 grammatical errors. They also pointed out that the errors in the aspect of the word were 1,946 i.e. 81.0% while in the sentence aspect, the number of offenses were 456 or 19.0%. Although the Malay language seems easy to speak or master, many still fail to use it according to the actual grammar standards (Norazlina, 2016). The main problem of Malay language is that many speakers have not been able to use the correct language when speaking even though most of them are local speakers including students at higher learning institutions (Mohd Juzaiddin et. al., 2006). Hence, the strengthening of Malay language at the local level should enhance from primary school students. The formulation of sentences describing the Malay texts has not gotten enough attention from researchers in Malaysia to the extent of building prototypes as it has with English (Yusnita and Zulikha, 2012).

In order to solve the problem of Malay Language grammar errors, the development of single sentence parser of Malay language was

developed. This parser receives an instruction from the user i.e. single sentence of Malay and classifies the sentence according to the grammar phrase. The built-in parser only receives instructions in the form of a single Malay sentence, classifies the sentence according to the grammatical phrases, and confirms the correct or incorrect sentences. The target group for this parser is primary school students.

## Related Works

This section focuses on research conducted by other researchers. There are five research papers selected to be discussed, namely Suzaimah Parser (Suzaimah, 2002), Juzaiddin Parser (Mohd Juzaiddin, 2006), Ahmad Parser (Ahmad, 2007), Hafhizah Parser (Noor Hafhizah, 2011), and BMTutor by Yusnita and Zulikha (2012).

### *Suzaimah Parser*

Suzaiman Parser was one of the Malay parsers that was built by one of the Universiti Putra Malaysia researchers (Suzaimah, 2002). This parser analysed sentences in syntax by using the top-down parsing method. This study did not involve the analysis of sentences in a semantic way to study the structure of the sentences included. Therefore, this parsing function was limited to only defining whether a sentence entered into the parsing system was valid or not.

This parser was built as a starting point because the parser was limited to the basic Malay texts such as single sentences and plural sentences using the parallel logical system function. This study used context-free grammar through the implementation of prologue clauses generated. This parser system was not tested on any real parallel computer. This is because the prologue language used was simply a simulation for logical parallelism.

The sentences entered in this parser would be interpreted by a natural language processing system to determine the validity of the sentence structure. The parsing of a given sentence depended on the grammatical structures formed through the prologue clause in this parser system. In conclusion, this study proved that prolog method is suitable for use in the development of natural language processing for Bahasa Melayu.

### *Juzaiddin Parser*

A group of researchers led by Mohd Juzaiddin Ab. Aziz (Juzaiddin et al, 2006) developed the parser. The parser used the Finite-State Automata (FSA) method and created a grammatical technique that did not require the lexical process to get a Part-Of-Speech (POS) sign for each processed word. Therefore, the purpose of this study was to classify the order of the sentences inserted using pattern techniques until a valid Malay sentence was established.

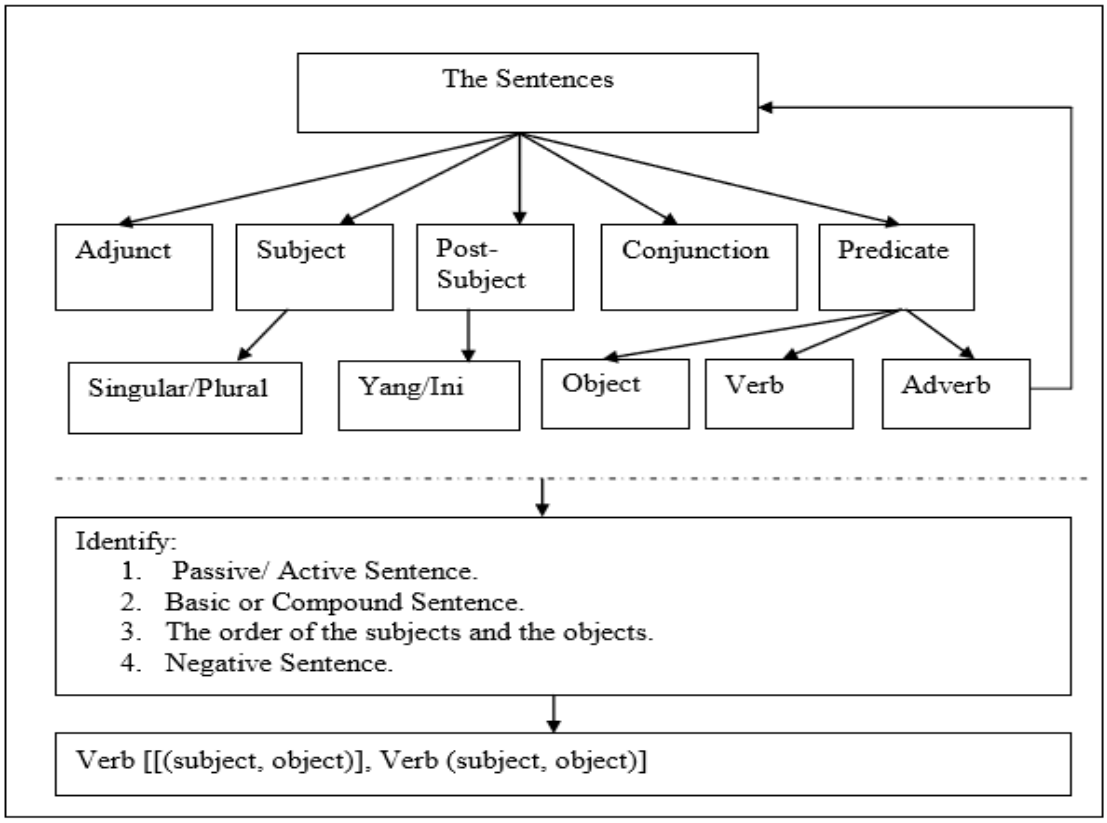


Figure 1: Juzaiddin Parser framework

***Ahmad Parser***

This parser was developed by a group of researchers from Petronas University Technology led by Ahmad Izuddin Zainal Abidin (Ahmad et al., 2007). This parser was a type of syntax parser using top-down parsing method. The purpose of this parser was to improve the existing parser by confirming the first sentence grammar. Another feature of this parser was to produce a parse tree based on the sentence entered if the sentence was grammatically valid

This parser used context-free grammar and prioritized the semantic part. In the semantic part, the Malay word was divided into two uses namely for human consumption and animal use. Some examples of special words for humans were *mengandung*, *memasak* and *memotong*. Meanwhile, examples of special words for animals were *meragut* and *bunting*. This parser could reduce the confusion of the use of words in the semantic part. Figure 2 shows the framework of this parser.

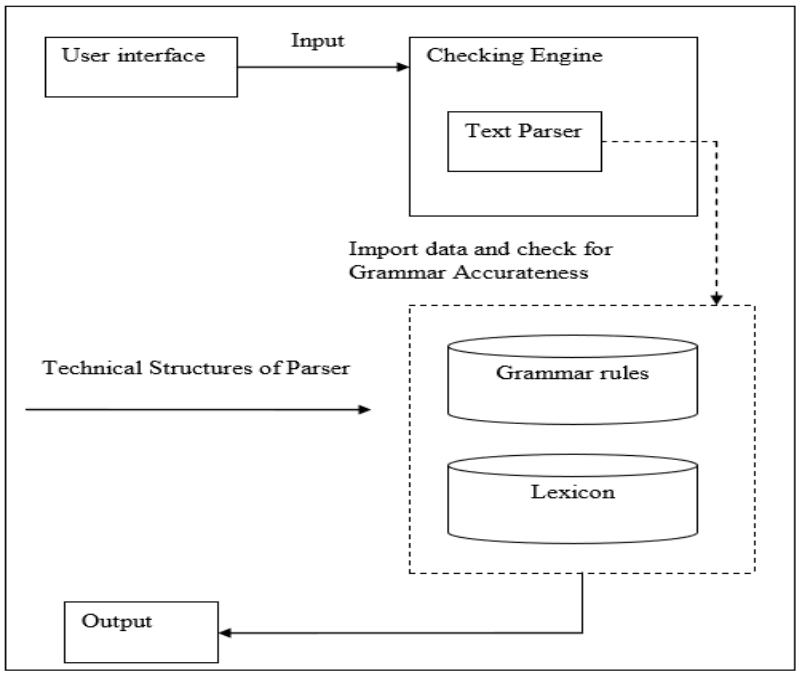


Figure 2: Ahmad Parser framework

### *Hafhizah Parser*

This parser was developed by one of the researchers from the University of Malaya, Noor Hafhizah Abd. Rahim, for her Master's degree (Noor Hafhizah, 2011). This parser was a syntactic type and used top-down decoding method. The purpose of this parser was to solve the problem of statistical parser, which was limited to Malay. In addition, this parser would generate a parse tree and would issue a parse tree, which recorded a high probability of value.

Figure 3 illustrates the framework for this parser. This parser used context-free grammar and emphasized the basic Malay verses. It improved the problems arising from synthesis of syntax using statistical method. This is because the syntactic method could not solve the confusion of the sentence structure. The statistical method used the probabilistic disclosure that applied to grammar in the decomposition. The advantages of this parser was in making the probable value calculation for each given sentence.

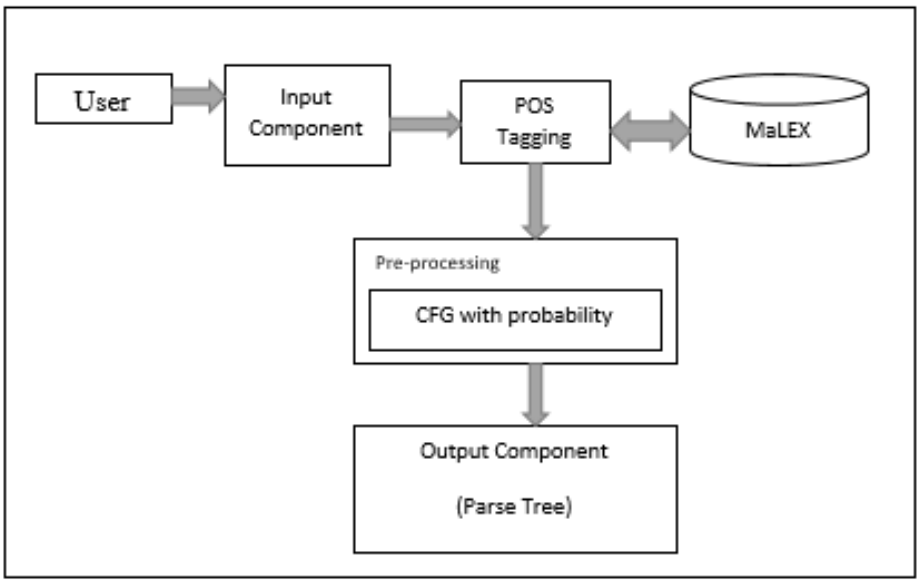


Figure 3: Hafhizah Parser framework

***BMTutor***

Two researchers from Universiti Utara Malaysia, Yusnita Muhamad Noor and Zulikha Jamaludin, developed BMTutor (Yusnita & Zulikha, 2012). This parser was syntactic and used context-free grammar. This parser used the token method by correcting Malay language errors using tokens, checking words, marking the Part-Of-Speech (POS) signs, confirming and matching the grammatical words described. Additionally, this parser would propose corrections and issue correct Malay words. This parser also worked to produce a parse tree and summarize the correct sentence based approximately on the sentence in the parse tree. Figure 4 shows the framework of this application.

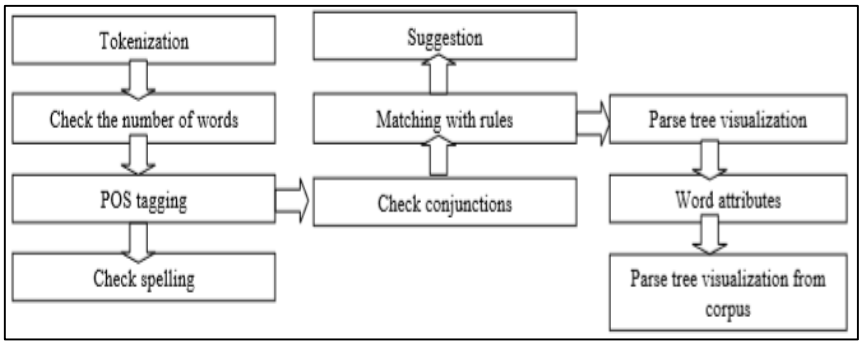


Figure 4: BMTutor Framework

**Methodology**

This section covers some of the research's processes and illustrates the framework of this Malay Single Sentence Parser. The framework of this study was designed to facilitate the planning of this research project in accordance with the designated plan. It is highly

recommended to avoid such things beyond the project's deadline. Figure 5 shows the framework of this study.

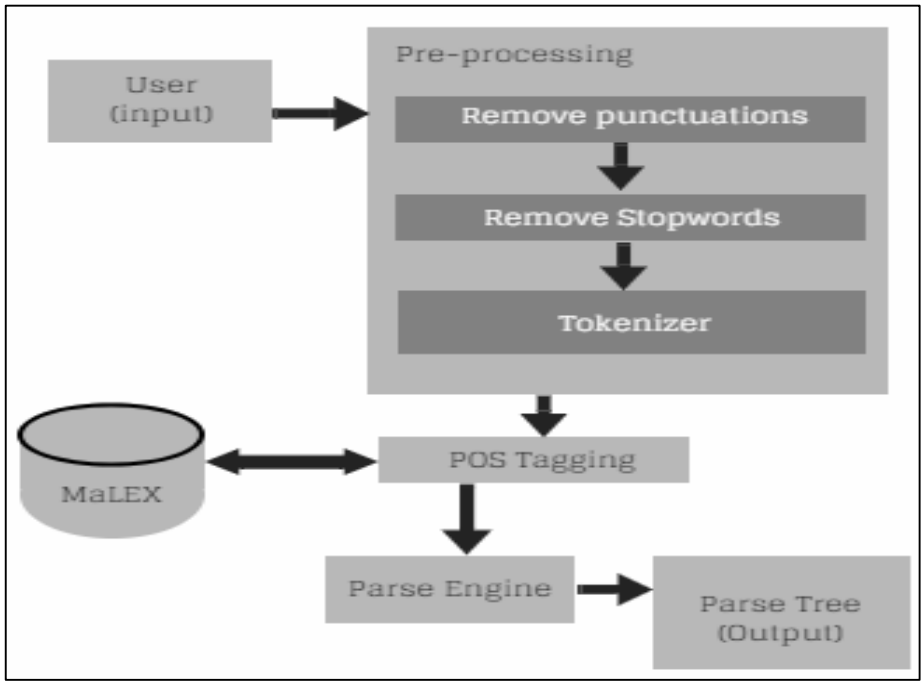


Figure 5: Malay Single Sentence Parser framework

User enters a single sentence into the parser. The input undergoes three parts of pre-processing where the first part is the removal of punctuations. This part removes all punctuations detected by the parser. Next, parser removes any stop words at the sentence such as *ialah*, *adalah* and *sedang*.

The final process is tokenization. Tokenizer splits the sentence into a few words. POS tagging process classifies the structure of the sentence referred to words in Malay Lexicon. Then, parser engine checks the structure of the sentence and displays the parse tree as output. The processes are illustrated in Figure 6 using an example of a sentence "*Ali sedang bermain bola*".

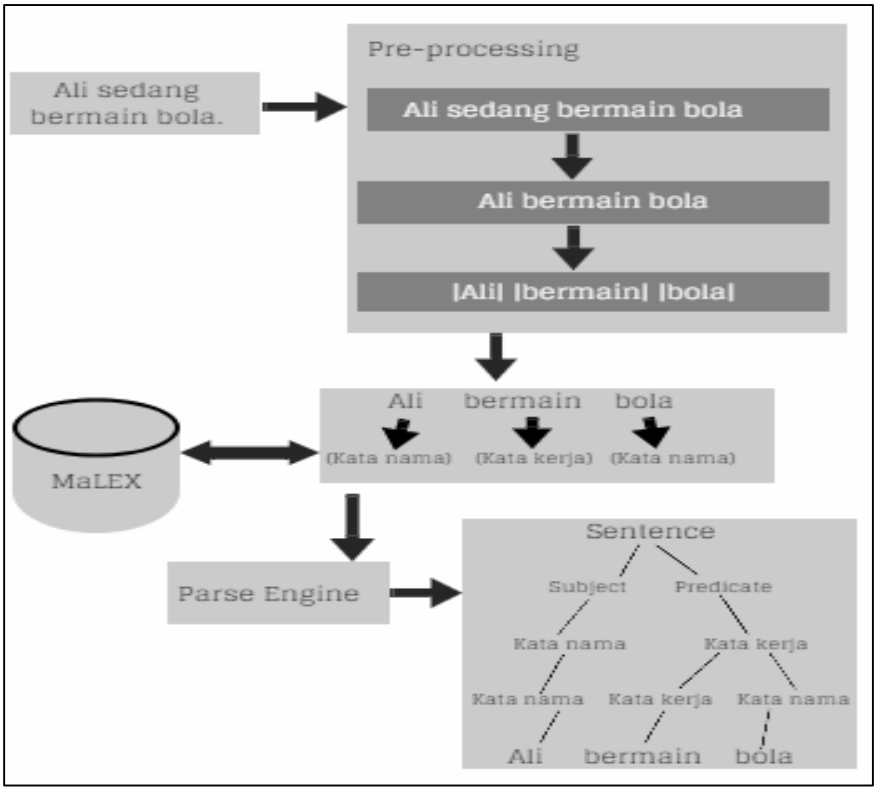


Figure 6: The processes in Malay single sentence parser

**Result and Discussion**

One of the main objectives of this study is to construct a single sentence of Malay. The evaluation to test the accuracy of the parser was done with three different metrics, namely recall, precision, and f-score (Carroll et al. 1998).

Precision technique was selected because this technique was best suited for this small sculpture study. While other techniques are very popular and have proved to be very accurate based on the results of previous studies, they were not appropriate for this study because the sample data was not suitable as a guide and there were time constraints in the evaluation of this parser.

A total of 41 single sentence samples were used to test the parser. The paragraph was derived from an expert, the Excellent Language Teacher, based on the citation of the *Dewan Masyarakat* magazine (Zahid, 2010). The parser recorded only 30 verses described correctly, while 11 more sentences were described as inaccurate by the parser. The precision of this parser was calculated using Precision technique and multiplied by 100. The parser recorded a percentage accuracy of 73.17%.

$$\text{The formula:} \quad Precision = \frac{Correct\_sentence}{Number\_of\_sample\_date} \times 100$$

$$\text{The result of experiments: } Precision = \frac{30}{41} \times 100 = 73.17\%$$

This accuracy was low as the parser only allowed the sentences entered based on excerpts only. Besides that, there was constraint in obtaining the Malay Language

Lexicon provided by *Dewan Bahasa dan Masyarakat*. Additionally, the parser only classified the subject according to the subject and the predicate based on the words. Sentences having multiple words in the subject area were considered inaccurate if the words entered did not meet the criteria. Furthermore, the parser was not able to distinguish between words used for humans and animals. For example, the word *ragut* has two different meanings; for humans it means steal while for animal it means eat.

## Conclusion

As a conclusion, a parser for Malay single sentence is presented. This parser is intended to help primary school students to understand the structure of a single Malay sentence. This parser serves to classify the structure of the sentence according to the correct phrases and remove it in the form of a parse tree easily understood. Students can identify the subject and predicate contained in a single sentence. In addition, this parser also serves to classify the special nouns found in the verse. However, this parser needs some improvements; for example, it should differentiate between the subjects and predicates without relying on words. In addition, this parser should also focus on parsing ambiguous words in order to distinguish the words used for humans and animals.

## References

Ahmad I. Z. Abidin, Yong, S. P., Rozana Kasbon & Hazreen Azman, (2007). Utilizing Top-Down Parsing Technique In The Development of a Malay Language Sentence Parser, Proceeding of the 2nd International Conference of Informatics, Universiti Malaya, Kuala Lumpur.

Baidura, K. & Jamilah, R. (2011). Analisis Kesalahan Tatabahasa Bahasa Melayu dalam Karangan Pelajar Asing di Sebuah Institusi Pengajian Tinggi Awam. Ms.Tesis. Universiti Islam Antarabangsa, Selangor.

Carroll, J. Briscoe, T. & Sanfilippo, A. (1998), Parser Evaluation: A Survey and a New Proposal, Proceedings of the First International Conference on Language resources and Evaluation, pp. 447-454.

Hamden, B. (2012). Morfologi dan sintaksis. Retrieved on 7 November, 2012, from http://hamdenbakar.blogspot.my/2012/11/morfologi-dan-sintaksis-1.html

Mohanty, S., & Balabantaray R. C. (2003), Intelligent Parsing In Natural Language Processing, 8th International Workshop of Parsing Technologies.

Mohd Juzaiddin, A.A. et. al., (2006). Pola Grammar Technique For GrammaticalRelation Extraction In Malay Language, Malaysian Journal of Computer Science, Vol 19, No. 1, pp. 59-72, University of Malaya.

Noor Hafhizah, A. R. (2011). A statistical parser to reduce structural ambiguity in Malay grammar rules, Ms Tesis. Universiti Malaya, Kuala Lumpur.

Norazlina, M.K. (2016). Bahasa Melayu Gagal Bahasa Inggeris Lemah. Retrieved on 29 February, 2016, from https://www.hmetro.com.my/node/118888.

Suzaimah, R. (2002). Reka bentuk dan implementasi suatu penghurai bahasa Melayu menggunakan sistem logik selari. Universiti Putra Malaysia, Selangor.

Yusnita, M. N. & Zulikha, J. (2012). Parser with sentence correction for Malay language (BM). Proceeding of International Conference on Information and Knowledge Management, 138-142.

Yusnita, M. N. & Zulikha, J. (2012). Malay Parse Tree Sentence Visualation (BMTUTOR). ARPN Journal of Engineering and Applied Sciences, 13116-13124.

Zahid M. Z. (2010). Kebanjiran Warga Asing Mendatangkan Keburukan. Retrieved on April 2010, from http://dwnmasyarakat.dbp.my/?p=48