World Applied Sciences Journal 12 (Special Issue on Computer Applications & Knowledge Management): 19-25, 2011
ISSN 1818-4952

# Malay Language Sentence Checker

*Rozana Kasbon, Nurul Atiqah Amran, Eliza Mazmee Mazlan and Saipunidzam Mahamad*

Computer and Information Sciences Department, Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia

**Abstract:** In the era of advanced technology, students are exposed to new gadgets such as hand phones and computer. Text messaging has brought a great deal of convenience and quickness to our daily communication. However, several problems have risen from the development and the used of the technology where many people tend to ignore the importance of grammar, spelling, syntax abbreviation and punctuation in their daily usage of texting, emailing and chatting. This has indirectly affects the students' ability when it comes to formal writings such as essay writing and report writing. This paper present the development of a Malay Language Sentence Checker, which translates short form texting language to its complete and correct forms, checks whether the sentences have any grammatical and structural errors and suggests the type of correction to the sentences. From the functional testing all the system components work as required. The user testing is also conducted and it is found that the system is more suitable to be used by students from lower secondary and primary schools.

**Key words:** Natural language processing · Malay language grammar · Parser

## INTRODUCTION

In linguistic, grammar is a set of logical and structural rules that govern the composition of sentences, phrases and words in any given natural language [1]. A sentence with a correct grammar representation allows people to communicate effectively in both written and verbal communication. Proper use of vocabulary, punctuation and standard grammar rules such as subject and verb allow communication to become more efficient. Grammar is like a road map where it guides us to arrive to the right destination. Without, a road map we might end up going to the wrong places at a longer time. In other words with a correct sentence structure we are able to convey our message clearly and understandable by others. The same affects may occur to non-native Malay speakers, where if they do not have adequate knowledge of Malay grammar it might cause communication problem and misunderstanding. Proper use of Malay language itself also applies to native speakers of Malay such as school and university students. This is might be due to the fact that they are not aware the proper use of Malay grammar since they have been talking and writing even though in Malay but with a mixture of other languages such as English and Chinese. As a result their written Malay is weak in which they are not able to write proper Malay sentence. In order to ensure the students regardless whether they are native Malay speakers or not, they need to master the Malay grammar.

On top of that, in the era of Information Technology, texting, chatting and emailing have been the most widely use medium of communication as the world is going for technological advancement. This medium of communication has caused people to take grammar for granted and they simply modify the language in both written and verbal communication. They have been using abbreviation and slang until they have forgotten how to develop proper sentences with a correct grammar representation. They need to have the knowledge on Malay Grammar since it will be an important key to an effective communication.

Grammar is also considered as natural language since it is a language that is spoken, written or signed by humans for general purpose communication. Natural Language Processing (NLP) which is a branch of artificial intelligent are designed to enable ordinary user to communicate with computer. Some of the applications of NLP include machine translation where it translates

**Corresponding Author:** Saipunidzam Mahamad, Computer and Information Sciences Department,
Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia.
E-mail: saipunidzam_mahamad@petronas.com.my.

human language text to another system such as database thus enabling human-language queries. The easiest tasks for NLP system are to parse a sentence to determine syntax; a more difficult task is determining the semantic meaning of a sentence and the most difficult task is the analysis of the context to determine the true meaning and comparing that with other text [2].

Currently, there are some systems that have been developed to detect and explain grammatical errors but most of them are for English sentences. In this paper we discuss the concept of developing a similar system in which the sentences that we usually use and type when we are chatting, texting or emailing will be converted to its longer form and will be checked with the Malay grammar. The system is purely using Malay language and is not a translator system. However, the developed system is well suited for both native and non-native Malay speaker. This system will guide the users on how to come up with proper sentence structure.

**Related Work:** A sentence that is produced by non-native Malay speakers who learn Malay Language might contain some errors if they do not know the proper grammar rules. For that reason they need to get some aid in order to learn Malay grammar. Since it is the era of technology, there is plenty of word processor software and tutorial system available to handle the problem [3]. However the product is usually produced for English language and other popular languages such as Arabic, Japan and Spanish. It is hard to find a grammar checker for Malay Language grammar and if the product exists it only focuses in text or word translator or as we known it, an e-dictionary.

Another example of a similar system available is GRADES, a diagnostic program that detects and explains grammatical mistakes made by non-native English Speaker [3]. The software is more efficient than other grammar checker since it perform diagnostic task, not through parsing but through application of classification and pattern matching rules. GRADES system is the generic tasked approach using hierarchical and hypothesis matching. There are two general classes of grammatical errors that are diagnose by GRADES; verb-related error which consist of the thirteen subtypes of error and non-related error which consist of four subtypes of error.

"An Expert System for the Teaching of English Grammar", is another example of language sentence parser that is written using Turbo Prolog Language. It makes use of context free, recursive descent parser to implement its grammar checking facility [4]. There are two environments available for the system, one for the teachers and one for the students. The teacher's environment allows the teacher to "teach" the system a variety of grammar that will be thought to the students where this is also called "user-defined" system. This structure needs the teachers to maintain its own set of grammar files to suit their needs. In the student's environment, the only function that the students can use is to check the grammar where it will display an error message if it detects any error in strings of word key in by the students.

Another example of an e-dictionary is Dewan Eja Pro as mentioned in [5]. It is a proofing and references suite for Malay language where it helps the users to write better sentences and makes them learn faster. It consists of spell checker and dictionaries endorsed by DBP. Its main function is the spell checker in which it is integrated with Microsoft Office where it automatically underlines errors and provides accurate suggestion. It is also integrated with an encyclopaedia where people can go from learning a new word to gaining new knowledge. The only limitation is that, it does not have grammar checker function.

A research was also being done by Yong *et al.,* in [6] that introduced a prototype of Malay sentence parser using top-down parsing technique. The prototype is only able to illustrate the structure of a grammatically correct sentence, determine if a sentence is grammatically and semantically correct.

There is also a system that is developed in [8] in which the system is used to steam out Arabic words that are used in Persian sentences.

Checking grammar can be generated using many techniques. Among the popular techniques are statistical-based checking and rule-based checking and syntax-based. In [9] a grammar checker is developed using a statistical-based approach in which n-gram words analysis and Part-Of-Speech (POS) tags are used for English and Bangla languages. The work in [10] describes a similar grammar checker that is developed for Punjabi language in which a rule-based checking is used. The rule-based technique is also using the POS tags that have all the grammatical information in which agreements checks at phrase and clause levels are done.

**Malay Grammar Structure:** The rules for Malay grammar structure were obtained from Tatabahasa Dewan [7]. There are four basic sentence patterns in Malay, arising from the way the predicate component is build up. All subjects are made up of the noun phrase (frasa nama), but the predicate component can be made up of any of the following phrases, namely Noun Phrase (Frasa Nama),

Table 1: Description of element used in Malay Grammar Phrase Structure

| Element | Description in Malay (English) |
|---|---|
| A | Ayat (Sentence) |
| Adj | Adjektif (Adjective) |
| AKomp | Ayat komplemen (Complementary sentence) |
| Bil | Bilangan (Numeric) |
| FA | Frasa Adjektif (Adjective Phrase) |
| FK | Frasa Kerja (Verb Phrase) |
| FN | Frasa Nama (Noun Phrase) |
| FS | Frasa Sendi ( Prepositional Phrase) |
| Gel | Gelaran (Title) |
| KB | Kata Bantu (Auxiliary) |
| Ket | Keterangan (Explanation) |
| KKtr | Kata Kerja Transitif (transitive verb) |
| KKttr | Kata Kerja Tak Transitif (Intransitive verb) |
| KNArah | Kata Nama Arah (Direction) |
| KNInt | Kata Nama Inti (Head noun) |
| KPeng | Kata Penguat (Intensifier) |
| Obj | Objek (Object) |
| P | Predikat (Predicate) |
| Pel | Pelengkap (Complement) |
| Pen | Penerang (Description) |
| PenjBil | Penjodoh Bilangan (Classifier)) |
| Pent | Penentu (Determiner) |
| S | Subjek (Subject) |
| SN | Sendi Nama (Preposition) |

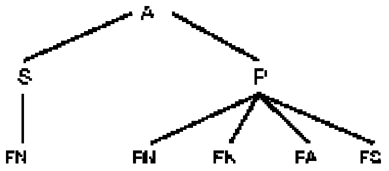


Fig. 1: Grammar Structure of a sentence

Verb Phrase (Frasa Kerja), Adjective Phrase (Frasa Adjektif) and Prepositional Phrase (Frasa Sendi Nama).

In this paper the Malay grammar structure uses the abbreviation shown in Table 1.

The term phrase is used in accordance with modern linguistic usage, in which constructing a sentence of at least one word have the potential of becoming bigger constructions. Fig. 1 shows the tree structure of the basic sentence. A sentence is comprises of subject (S) and a predicate (P) where S is a noun phrase (FN) and P may consist of a combination of noun phrase (FN), verb phrase (FK), adjective phrase (FA) and preposition phrase (FS).

A noun phrase (FN) is made up of a word or a series of words, of which a noun is the major item. It is actually a complex structure and made up of the following component which is shown in the tree structure in Fig. 2.

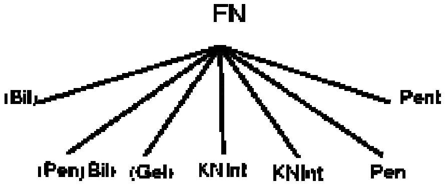


Fig. 2: Noun Phrase Grammar Structure

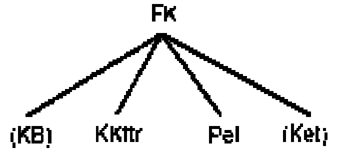


Fig. 3: Intransitive Verb Phrase Grammar Structure

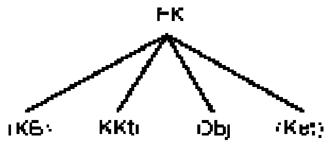


Fig. 4: Transitive Verb Phrase Grammar Structure

A verb phrase (FK) may also occupy the position of predicate. A verb phrase may consist of a single verb or a series of words one of which is either a transitive or an intransitive verb. The verb in Malay sentence is the word which is central in a verb phrase. Intransitive verb are verbs which do not require objects. There are two categories of intransitive verb:

- Intransitive verb without complement, where the verb occurring alone in their own.
- Intransitive verb with complement, where it accompanied by additional elements necessary to complete the meaning of the verbs.

Fig. 3 shows the grammar structure of intransitive verbs where it consists of KB (auxiliary), KKttr (intransitive verb), Pel (complement) and Ket (explanation).

Transitive verbs are verb which require objects. Objects following transitive verbs are made up of noun phrases. Transitive verbs can be an active sentence or a passive sentence. Figure 4 shows the grammar structure of intransitive verbs where it consists of KB (auxiliary), KKtr (transitive verb), Obj (object) and Ket (explaination).

The adjective phrase (FA) is a series of words of which the main element is an adjective. An adjective may occur on its own or with more or one intensifiers. There

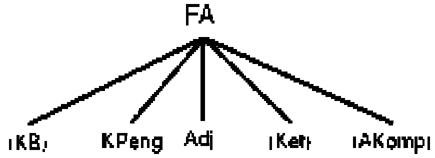


Fig. 5: Adjective Phrase Grammar Structure

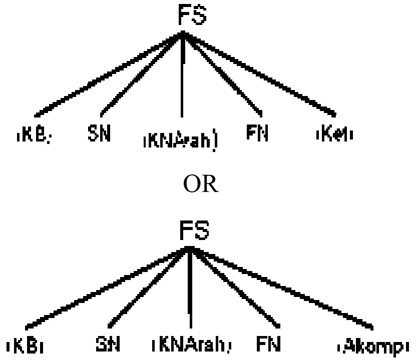


Fig. 6: Prepositional Phrase Grammar Structure

are three types of intensifiers; it can be in front of adjectives, after adjectives or it can occur either in front of or after adjectives. The adjectives can also be accompanies by the auxiliary. Figure 5 shows the grammar structure of adjective phrase where it consists of KB (auxiliary), KPeng (intensifier), Adj (adjective), Ket (explaination) and Akomp (complementary sentence).

The prepositional phrase (FS) is a construction of which the main elements are a preposition followed by a noun phrase. Sometimes a prepositional phrase may be formed with a direction word inserted between the preposition and the noun or noun phrase. Some prepositional phrase consists of a preposition followed by a noun phrase and a modifier. Figure 6 shows the grammar structure of preposition phrase where it can contain KB (auxiliary), SN (preposition), KNArah (direction), FN (noun phrase), ket (explaination) or Akomp (complementary sentence).

**System Design:** The development of this project is divided into three different modules which are, Module 1: Translating short form texting sentence to its longer and complete form, Module 2: Categorizing the sentence based on its lexicon, Module 3: Checking the sentence with Malay grammar structure.

In Module 1 the short form sentence is translated into its longer and complete form for example the word '*mkn nsi*' will be translated to '*makan nasi'*. The steps taken are as follows:

- Split user input
- Check the input with abbreviation database
- Display longer form
- Combine the sentence
- Display output

In module 2, the words are categorized based on the word category in the database. There are 57 categories of word currently available in the database. The examples of the categories are *Kata Nama Am, Kata Nama Khas, Kata Adjektif* and many more. The steps taken are as follow:

- Split long form sentence
- Check the sentence with lexicon database
- Get the word category name
- Display split sentence and its category

**Example of the Input:** *'Kami membuat rumah untuk ibu'*

**Example of the Output:**

- *Kami-Kata Ganti Nama Diri Orang*
- *Membuat-Kata Kerja Transitif Aktif*
- *Rumah-Kata Nama Am-Konkrit*
- *Untuk-Kata Sendi*
- *Ibu-Kata Nama Am-Manusia*

Module 3 checks the grammar structure with the rules discussed in section 3 (Malat Grammer Structure). Rule-based checking technique is used to check the grammar structure in a sentence. The steps are as follows:

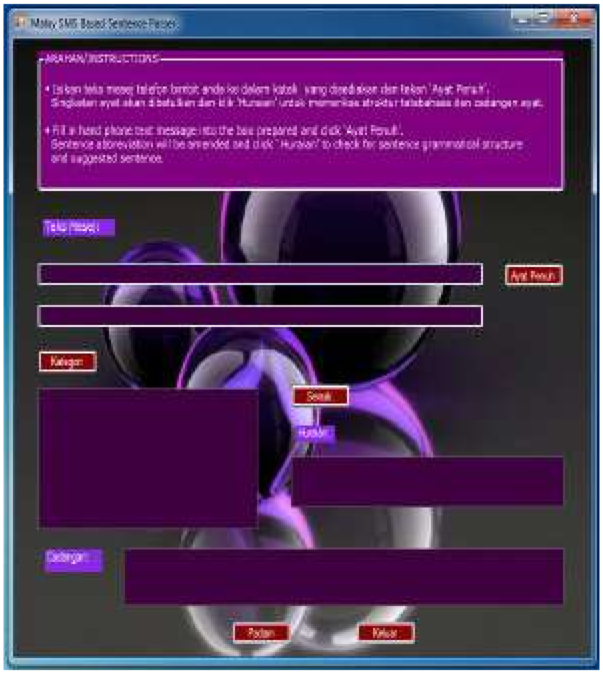
- Get the sentence from module 2.
- Check the category of sentence with the Malay Grammar structure rules. For eg. for a sentence to be categorized as correct *FA*, it may consist of elements on the right-hand side of their respective rules or combination of the rules.

*FA _ (KB) + (KPeng) + Adj + (Ket) + (AKomp)*

**Display the Output of the Sentence**
**Suggest the Format of the Sentence**
**Example of the Input:** *'Kami membuat rumah untuk ibu'*

**Example of the Output:**

- *Ayat ini mengandungi kesalahan.*
- *Ayat ini harus mengandungi penjodoh_bilangan.*


Malay SMS Based Sentence Parser

ARAHAN/INSTRUCTIONS

- Isikan teks mesej telefon bimbit anda ke dalam kotak yang disediakan dan tekan 'Ayat Penuh'. Singkatan ayat akan dibetulkan dan klik 'Huraian' untuk memeriksa struktur tatabahasa dan cadangan ayat.
- Fill in hand phone text message into the box prepared and click 'Ayat Penuh'. Sentence abbreviation will be amended and click 'Huraian' to check for sentence grammatical structure and suggested sentence.

Fig. 7: User Interface

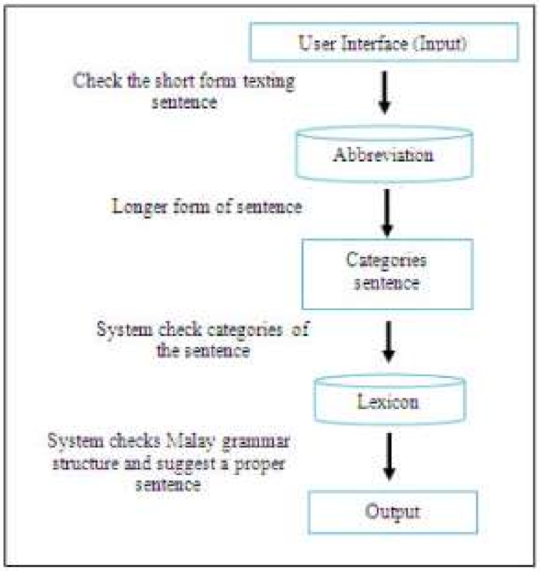

User Interface (Input)
Check the short form texting sentence
Abbreviation
Longer form of sentence
Categories sentence
System check categories of the sentence
Lexicon
System checks Malay grammar structure and suggest a proper sentence
Output

Fig. 8: System Architecture

All three modules are performed using a single interface as shown in Fig. 7.

Friendly user interface design allows the users to experience the system better. User interface provides a systematic view on how the users should navigate the system. Instructions are placed inside the box on top of the page. User needs to enter the sentence into first textbox then click 'Ayat Penuh' button. Next, user can click 'Kategori' to categories the sentence. Lastly user can click 'Semak' to check for grammar structure.

System architecture is as shown in Fig. 8 is the eagle view of the system prototype, how the database and interface design are connected and the process or sequence of step that produce the system prototype within constraint and requirement. Lexicon is a set of available words in a given context. Every word can be classified through a lexical category such as noun, verb, adjective, adverb, conjunction, preposition or pronoun.

## RESULTS AND DISCUSSION

Functional testing is done for the system to detect any errors. Table 2 shows the component available in the system interface. Each component listed is functioning as expected.

In order to test the accuracy of the system, user evaluation is also conducted. The test users are teachers and students of a tuition center, a primary school, a secondary school and a university.

There are 45 people involved in this evaluation. Each of them is given with 2 to 6 samples of testing sentence. They are randomly selected from different race, age and educational background.

A total N of 170 grammatically correct sentences was provided by those students as the test cases the prototype. The average result Avg equation (1) and the weighted average result WAvg equation (2) of correctly checked sentences were calculated using the following two equations:

$$Avg = \frac{100\% \times \sum_{i=1}^{k} \frac{B_i}{A_i}}{k} \tag{1}$$

where $k$ = 4 representing four school levels, $A$ is the number of test cases and $B$ is the total number of sentences correctly checked.

$$WAvg = 100\% \times \sum_{i=1}^{k} \frac{A_i}{N} \times \frac{B_i}{A_i} \tag{2}$$

where $k$ = 4 representing four school levels, $N$ represents the total number of test cases, $A$ is the number of test cases and $B$ is the total number of sentences correctly checked. The weighted average is calculated mainly because the number of test cases was different for the four school levels. The results are shown in Table 3.

Table 2: Functional Testing for Each System Component

| Component | Expected Test Result | Testing Result |
|---|---|---|
| Ayat Penuh Button<br>-Get user input<br>-Split sentence<br>-Check with abbreviation database | Translate short form texting to longer form | Able to capture user input, split and translate |
| Kategori Button<br>-Split sentence<br>-Check with lexicon database | Categories sentence based on lexicon | Able to categories sentence |
| Semak Button<br>-Check Malay grammar<br>-Suggest sentence format | Produce sentence grammar result | Able to check and suggest sentence with correct Malay grammar |
| Padam Button<br>-Erase all textbox and labels | Empty all textbox and label | Able to empty all textbox and label |
| Keluar Button | Ensure the system is closed | The system successfully closed |

Table 3: Results of test cases

| School Level | Test Cases Provided (A) | Total Sentences Correctly Checked (B) | Result: (B/A) x 100% |
|---|---|---|---|
| Primary | 27 | 24 | 88.9 % |
| Tuition | 54 | 45 | 83.3 % |
| Secondary | 27 | 21 | 77.8 % |
| University | 62 | 48 | 77.4% |
| | Average | | 81.9 % |
| | Weighted Average | | 80.6 % |

The testing results show that the prototype was able to achieve the average rate of 81.9% from the sentences parse to the prototype. The prototype scored higher for sentences provided by primary and tuition students while lower score was obtained from sentences given by the secondary school and university students. This might be due to the fact the prototype is more suitable for students who are in the lower secondary schools and primary schools because the Malay language use at secondary and university levels is more complex than what the prototype can offer.

## CONCLUSION

The developed system is able to translate short form sentences to its longer and complete form, then categorizing the sentences based on its lexicon and finally checking the grammar structure and suggesting the correction to be made to the sentences. The result of functional testing proves that all the system components work as required in coming up with final suggestion to correct the grammatical errors in the given sentences. Even though from the user testing that it can be concluded that the prototype is suitable to be used by the primary school students, other level of students might also benefit from it. It is also suited for both native and non-native Malay speakers.

## REFERENCES

1. Wikipedia the free encyclopedia, Retrieved February 2010, from the World Wide Web: http://en.wikipedia.org/wiki/Grammar.
2. Terry, H., 2010. PCAI-Where Intelligent Technology Meets the Real World, Retrieved February 2010 from the World Wide Web: http://www.pcai.com/web/ai_info/natural_lang_proc.html.
3. Fox, R. and M. Bowden, 2002. Automated Diagnosis of Non-Native English Speaker's Natural Language. The Tools with Artificial Intelligent (ICTAI 2002) 14th IEEE International Conference, Washington DC, USA, 4-6 November 2002, pp: 310-306.
4. Keong, C.C., 1990. An Expert System for Teaching of English Grammar. 1990 IEEE Region 10 Conference on Computer and Communication Systems TENCON 90, Hong Kong, China, 24-27 September 1990, pp: 722-727.
5. Dewan Eja Pro, Retrieved February 2010, from World Wide Web: http://www.tntsb.com/index.php?q=DewanEjaPro.
6. Yong, S.P., A.I. Zainal Abidin and R. Kasbon, 2007. Utilizing Top-Down Approach in Malay Language Parsing System. 2$^{nd}$ International Conference on Informatics, Kuala Lumpur, Malaysia, 27-28 November 2007, pp: 56-60.

7. Karim, N.S., F.M. Onn, H. Musa, A.B. Mahmood, 2009. Tatabahasa Dewan Edisi Ketiga, Dewan Bahasa dan Pustaka.
8. Yoosofan, A., A. Rahimi, M. Rastgoo and M.M. Mojiri, 2010. Automatic Stemming of Some Arabic Words Used in Persian through Morphological Analysis without a Dictionary. World Appl. Sci. J., 8(9): 1078-1085.
9. Alam, M.J., N. UzZaman and M. Khan, 2006. N-gram based Statistical Grammar Checker for Bangla and English. 9th International Conference on Computer and Information Technology (ICCIT 2006), Dhaka, Bangladesh, 21-23 December 2006, pp: 119-122.
10. Gill, M.S. and G.S. Lehal, 2008. A grammar Checking System for Punjabi. 22nd International Conference on on Computational Linguistics: Demonstration Papers (COLING '08), Manchester, UK, 18-22 August 2008, pp: 149-152.