POLA GRAMMAR TECHNIQUE FOR GRAMMATICAL RELATION EXTRACTION IN MALAY LANGUAGE

***Mohd Juzaiddin Ab Aziz*[1], *Fatimah Ahmad*[2], *Abdul Azim Abdul Ghani*[2], *Ramlan Mahmod*[2]**

[1]Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 Bangi,
Selangor, Malaysia. Email: din@ftsm.ukm.my
[2]Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Putra Malaysia, 43400 Serdang,
Selangor, Malaysia.
Email: {fatimah, azim, ramlan}@fsktm.upm.edu.my

ABSTRACT

A basic sentence in Malay language is either a combination of NP+NP, NP+VP, NP+PP, or NP+AP. The language is a structure phrase grammar. The Context Free Grammar was developed by Nik Safiah (1993). However, in order to derive a parse tree for a syntactic process, the CFG was found to be complicated due to many ambiguities for part of speech. This paper shall introduce a pola-grammar technique that does not require lexical process of retrieving the part of speech for each word. The techniques used are the automata and the finite states. During the process, sentences will be grouped into an adjunct, a subject, a post-subject, a conjunction and a predicate. The predicate consists of a verb, a conjunction, an object and an adverb. An adverb consists of a verb, a conjunction and an adverb. These components which are in sequence order are called pola. The subject, the object and the verb can be identified by making use of the pola in the sentence. This can be done by spliting the sentence into its pola. The information in the predicate will be processed to get the verb and object.

Keywords : Malay language, Grammar, Pola Grammar, Grammatical Relation.

1.0 INTRODUCTION

The specific task discuss in this paper is to identify the grammatical relations such as subject, object and adjunct in the Malay language. The information is very useful for computer applications such as an application to annotate the thematic roles in semantic analysis. The data used in this experiment was taken from the collections of Computer Science and Information Technology thesis's abstract in Perpustakaan Tun Seri Lanang, Universiti Kebangsaan, Malaysia.

2.0 MALAY LANGUAGE'S GRAMMAR

According to the research done by Azhar [2], there are three types of Malay language's grammar. First, sentence grammar developed by Nik Safiah [8] and Yeoh [17], second, partial discourse grammar Asmah [13] and third, 'pola' grammar.

2.1 Sentence Grammar

The sentence grammar is based on the models of transformation-generative grammar and the relational grammar of English [2]. The Malaysia ayat grammar developed were inherently from the phrase structure grammar (PS-grammar) that was developed by N. Chomsky in 1957 [8][17]. The PS-grammar which restricts on the form of PS-rules result in regular, context-free grammar (CFG) , context-sensitive, and unrestrictred PS-grammar.

A CFG is called a context-free because the left-hand side of a type 2 rule consists by definition of a single variable. To verify the syntax, a CFG for Malay language was developed by Nik Safiah [9] and it is shown in Fig. 1. The figure shows that the terminals for the language are Ayat, Subjek, Predikat, Frasa Nama, Frasa Kerja, Frasa Adjektif, and Frasa Sendi.

Ayat	→ Subjek + Predikat
Subjek	→ Frasa Nama
Predikat	→ Frasa Nama | Frasa Kerja| Frasa Adjektif| Frasa Sendi
Frasa Nama	→ (Bil) (Penj Bil) (Gelaran) Kata Nama Int <Kata Nama Int> (Penentu) (Pent)
Frasa Kerja	→ (Kata Bantu) [KKtr (obj | Akomp) | KKttr (Pel | Akomp)] (Ket)
Frasa Adjektif	→ (Kata Bantu) (Kata Penguat) + Adj + (Ket) + (Akomp)
Frasa Sendi	→ (Kata Bantu) + Sendi Nama + (Kata Nama Arah) + FN + (Akomp | Ket)

Fig. 1: A context free grammar for Malay language

A parse tree is important in order to derive the grammar with CFG. It shows the validity of phrases exists in the sentence. But, the main problem in deriving the parse tree is the ambiguity problem. For example, the parse trees in Fig. 1 and 2 show the derivation for the basic sentences, (1) and (2).

(1) Dia menulis aturcara.
(2) Dia gemar melancong.
(3) Dia gemar menulis aturcara.

Both sentences (1) and (2) do not have any grammatical error according to the CFG. Let, insert the verb “gemar” in sentence (2) into sentence (1), and produce a new sentence (3) which also does not have any grammatical error. But, in sentence (3), the word “gemar” can either be a “verb” or an “auxilliary” (kata bantu). If it is recognized as a ‘verb’ then the sentence can not be parsed by the CFG, even though the sentence is valid. A new parse tree has to be reconstructed and the word is recognized as an “auxilliary” to produce a valid result. In this case, it is called a word ambiguity where a word is ambiguous if they hold more than one part-of-speech (POS). The parse tree for sentence (3) is shown in Fig. 4.

Another type of ambiguity problem occurs when more than one parse tree are constructed. In this case, the algorithm has to decide which tree is the true valid tree. The invalid production will cause a wrong process or an ill-grammar problem.

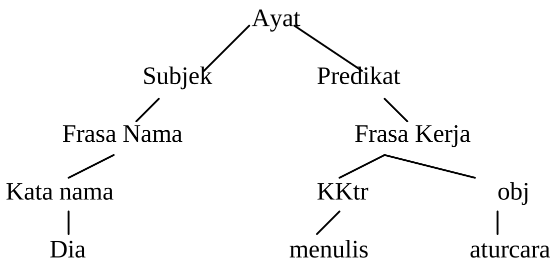

Fig. 2: A derivation of sentence (1)

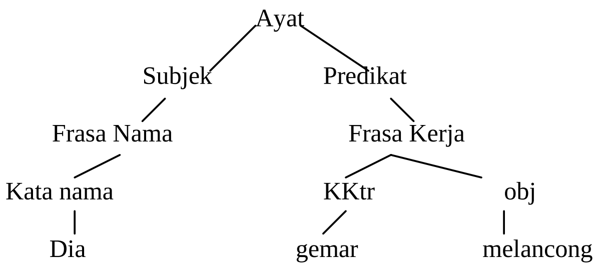

Fig. 3: A derivation of sentence (2)

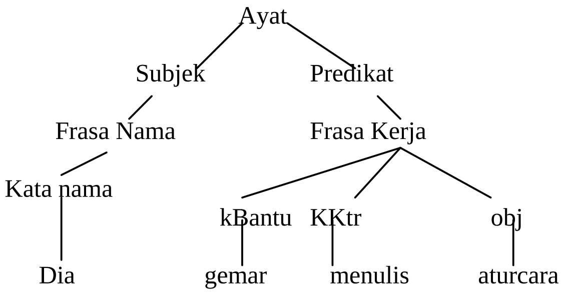

Fig. 4: A derivation of sentence (3)

2.2 Partial Discourse Grammar

A discourse grammar is a grammar of sentences used in discourse. It is the grammar that picks out sentences from discourse in order to make linguistic statements about them. A sentence grammar uses the theory-first approach, while the partial discourse grammar uses the language first approach in the writing of syntax [2].

2.3 The Pola Grammar

The term pola refers to 'pattern' as in "sentence pattern". Pola grammar is the term used by Azhar [2] as a cover term for work describing Malaysia grammars. A pola grammar is mainly a list of "pola-pola ayat" or sentence patterns. Each pola is the concatenation of the class-names of sentence constituents. Each pola is a formula for making one type of basic sentences.

In the Malay language as described by Lewis [4], the 'pola' or patern has four elements: the Measurer, the Head, the Qualifier and the Determinant. The work is affirmed by the elements of Malay language structure discussed in [10][7][14]. Asmah [3] and Abdullah [1] proved that the Malay language is constructed from a 'pola sentence' or sentence patern. The following are the pola grammar of the basic sentences, taken from Asmah [3].

i. Pelaku + Perbuatan (Actor + Verb)
ii. Pelaku+Perbuatan+Pelengkap (Actor + Verb + Complement)
iii. Perbuatan + Pelengkap (Verb + Complement)
iv. Diterangkan + Menerangkan (Signified + Signified)
v. Digolong + Penggolong (Classified + Classifier)
vi. Pelengkap + Perbuatan + Pelaku (Complement + Verb + Actor)
vii. Pelengkap + Perbuatan (Complement + Verb)

3.0 SENTENCE

There are two types of Malay language's sentence [10][15]. The first type is the basic sentence and the second type is the compound sentence. A basic sentence consists of a subject and an object while a compound sentence can either has:

i. A subject with more than one object.
ii. More that one subjects with more that one object.
iii. More that one subject with one object.

Example of a basic sentence and compound sentence are shown in sentence (4), (5) and (6). Sentence (4) is a basic sentence consists of one subject and one object. Sentence (5) consists of one subject and two objects while sentence (6) consists of two subjects. The first subject in (6) is "pengkompil" and the second subject is "penterjemah". The objects for "pengkompil" are "leksikal", "sintaksis" and "semantik". The object for "penterjemah" is "aturcara". Sentence (7) consists of three subjects and one object.

(4) Tugas pengkompil adalah menterjemah aturacara.
(5) Tugas pengkompil adalah menterjemah aturcara dan menyemak sintaks.

(6) Tugas pengkompil adalah melakukan analisis leksikal, sintaksis, dan semantik, dan tugas penterjemah menterjemah aturcara.
(7) Penganalisis Leksikal, Penganalisis Sintaks dan Penjana Kod Pertengahan merupakan fasa yang berlainan dalam pengkompil.

4.0 PARSING AND ITS PROBLEM

A variety of different approaches have been taken for robust extraction of grammatical relations. Dependency parsing is a natural technique to use, and there are some works in that area on robust analysis and disambiguation [12].

Many languages use sentence grammar or a CFG to validate the grammar after the Chomskyan revolution. But, ambiguity is still a major problem in parsing with a sentence grammar [11]. For example, a part-of-speech (POS) ambiguity, shown in sentence (8). In the sentence, the phrase “menulis aturcara” is a Verb Phrase but it appears as the header of a sentence. Based on the CFG given in Fig. 1, the sentence is a subject+predicate, and a subject is a noun phrase. In the case of the sentence (8), it is valid in two conditions:

a. a verb phrase can be the subject.
b. “menulis” is a noun.

(8) Menulis aturcara merupakan kegemarannya.

Sentence (8) can be rewritten in the formation of noun phrase + verb phrase, as shown in the sentences, (9). Sentence (10), is another sentence that begins with a verb phrase, and holds the same meaning with sentence (8).

(9) Dia gemar menulis aturcara.
(10) Kegemarannya adalah menulis aturcara.

The sentence (8) and (10) are not ill-grammar sentences. So, the first phrase appears in the sentence might be a noun phrase where the words “menulis” and “kegemarannya” have ambiguity values of POS. Assume that the first segment of the sentence is the subject, then the subject for sentence (8) is ‘Menulis aturcara” and the predicate for sentence (8) is ‘merupakan kegemarannya’. ‘kegemarannya’ will be the object for the sentence (8).

For sentence (9), the subject is ‘Dia’ and the predicate is ‘gemar menulis aturcara’. The object is ‘menulis aturcara’. In sentence (10), the subject is ‘Kegemarannya’, the predicate is ‘adalah menulis aturcara’ and the object is ‘menulis aturcara’. Table 1 shows the result of the subject, predicate and object mentioned above.

Table 1: The pola subject, predicate, object for sentences, (4), (5) and (6)

Sentence	Subject	Predicate	Object
(8)	Menulis aturcara	Merupakan kegemarannya	Kegemarannya
(9)	Dia	Gemar menulis aturcara	Menulis aturcara
(10)	Kegemarannya	Adalah menulis aturcara	Menulis aturcara

The result in Table 1 shows that the subject for sentence (8) is equivalent to the objects in sentences, (9) and (10). The verbs for the sentences are:

a. “merupakan”, for sentence (8)
b. “gemar”, for sentence (9)
c. “menulis” for sentence (10)

5.0 POLA GRAMMAR TECHNIQUE

In the Malay language, an alternative technique to CFG is the pola grammar technique. The basic pola grammar is the pola which were produced by Asmah [3] and Nik Safiah [9]. The pola are the structure phrase of:

i. Noun Phrase + Noun Phrase

ii. Noun Phrase + Verb Phrase
iii. Noun Phrase + Preposition Phrase
iv. Noun Phrase +Adjective Phrase

The structure phrase above will be our basic guide to produce the five pola:

i. adjunct
ii. subject
iii. post-subject
iv. conjunction
v. predicate

The examples of using the pola grammar above are shown in sentence (11), (12) and (13).

(11) Pengkompil menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah.
(12) Tujuan pengkompil adalah untuk menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah.
(13) Walaupun pengkompil menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah, tetapi, tugas utamanya adalah untuk menyemak sintaks bahasa.

In sentence (12), the adjunct ‘Tujuan’ is added to the sentence (11). The word ‘adalah’ is refered as the pemeri (affirmation) to the nouns ‘Pengkompil’ and the word ‘untuk’ is the prepositions for the subject and predicate. ‘Walaupun’ appear in sentence (13) which is a conjoined sentence. The sentence is formed from two simple sentences through a conjunction, ‘tetapi’.

The pola in the sentence appear in sequence of adjunct, subject, post-subject, conjunction, and predicate. In order to construct the algorithm to identify each pola in a sentence, the subsections below discuss in details each of the pola mentioned above.

5.1 The Adjunct

An adjunct is a type of adverbial illustrating the circumstances of the action. It appears as the first segment in the sentence. Examples of the adjuncts are listed in Table 2. The properties are:

a. Type : Numeral, numeral classifier and subordinating conjunction.
b. Role : To support the subject
c. Other : Can be followed by the word ‘yang’ to describe further of the type.

Table 2: An example of the adjuncts

Numeral	**Numeral Classifier**	**Numeral**	**subordinating / conjuction**
Dua (numeral) Pada Di	Orang Beberapa Puluh	Lampau Silam Orang	Kerana Agar Sekiranya

5.2 The Subject

The subject tells whom or what the sentence is all about. The subject of a sentence contains the ‘who’ or what described by or acting in the sentence. The example of the subjects, post-subject, conjunction and predicate are shown in Table 3. The properties of the subjects are:

a. Type : Noun, Pronoun, Verb (that behave as noun).
b. Role : To tells whom or what the sentence is all about.
c. Other : Can be followed by the word ‘yang’ to describe further of the type.

Table 3: An example of segment in the sentences

Subject	Post-subject	Conjunction	Predicate
Pengkompil		Adalah untuk	Menterjemah aturcara.
Pembolehubah	Yang berjenis integer		Tidak boleh diistiharkan di kawasan tersebut.
Penganalisis Leksikal, Penganalisis Sintaks dan Penjana Kod Pertengahan		Telah	Dibangunkan dalam aturcara tersebut.
Pengkompil,	Berbeza dengan penterjemah,		Melaksanakan tugasnya setelah mendapat input secara pukal.

5.3 Post Subject (Yang)

The word ‘yang’ is normally used in Malay language because the Malay language is a terse language. Some of the nominal and verbal information may not be available within a clause but available only from its context. The example is in Table 2, in the second column. The properties are:

a. Type : ‘yang’, coma,
b. Role : intensifier.

Problem arises when marking the end of the clause that begin with ‘yang’. This research identifies the end of the clause by using the coma sign, the modifier, ‘ini’ and ‘itu’ and the conjunction. Anyway, not all “yang’s” clause will stop at these markers.

5.4 Conjunction

Expressions that may occur in the sentences are usually from the groups of conjunction, adverbials and prepositions [5]. However, this paper will assume all of these terms as a conjunction. The definition of the conjunction by Nathesan [6], is the word used as discourse marker or “kata penghubung”, to show a continuity of idea.

5.5 The Predicate

The predicate tells something about the subject. A simple predicate is the predicate verb, being verb, or verb phrase associated with the sentence’s subject. A complete predicate is a simple predicate and any modifier. Predicate modifier may be adverbs, and adjectives.

6.0 CONCEPTUAL DESIGN

In order to test the pola grammar automatically, a conceptual design of the grammar was developed. The design contains 30 rules for the pola grammar. It begins with the pola for the basic sentence,

NP+NP (1)
NP+VP (2)
NP+PP (3)
NP+AP (4)

The basic sentence consists of

Subject (NP) + Predicate (NP) (5a)
OR

Subject (NP+NP) + Predicate (Null) (5b)
Subject (NP) + Predicate (VP) (6)
Subject (NP) + Predicate (PP) (7)
Subject (NP) + Predicate (AP) (8)

A subject is a Noun Phrases and a predicate consists of,

From (4a), Predicate (N + Adjective) (9)
From (5), Predicate (V + Adverb) (10a)
OR
From (5), Predicate (V) (10b)
OR
From (5), Predicate (V + Object) (10c)
OR
From (5), Predicate (V + Agent) (10d)
From (6), Predicate (P + Object) (11)
OR
From (6), Predicate (P + Agent) (11b)
From (7), Predicate (A + Agent) (12)

An adjuncts can be included in a sentence as shown below,

Adjuncts + NP+NP
NP+VP
NP+PP
NP+AP

The rules requires that, if there are adjuncts,
From (4a), Adjuncts + subject (N) + postSubject + Predicate [conjunction + X],
where X = NP or Null. (13)

From (5), Adjuncts + subject (N) + postSubject + Predicate[conjunction + Y],
where Y = V or V + Adverb or V + Object or V+Agent (14)

From (6), Adjuncts + subject (N) + postSubject + Predicate [conjunction + Z],
where Z = Object (15)

From (7), Adjuncts + subject (N) + postSubject + Predicate [conjunction + N].
where N = Object (16)

For the passive sentences, the rules are,

Subject (NP) + VP [di + verb oleh Object] (17)
Subject (NP) + VP [noun + Verb] (18)

The rules for the compound sentence are,

First,
Subject + Predicate + Conjunction + subject + predicate
From (1) to (11)

NP+NP		Conjunction		NP+NP	(19)
NP+VP				NP+VP	(20)
NP+PP	+		+	NP+PP	(21)
NP+AP				NP+AP	(22)

Second,
Subject + predicate
From (1) to (12)

NP	+	Conjunction	+	NP VP PP AP	(23) (24) (25) (26)

Lastly,

Subject + Predicate+ Conjunction + predicate
From (1) to (11)

NP+NP NP+VP NP+PP NP+AP	+	Conjunction	+	NP VP PP AP	(27) (28) (29) (30)

6.1 Theoretical Prove

The output of the process is a Dependent Structure (verb which related to the root) of a sentence. The format will be,
Verb (subject, Objects)

a. First, to test the design, take the sentence (8) as the input.
"Pengkompil menukar bahasa paras tinggi kepada bahasa mesin".

→ Step 1

Identify, it is a Basic Sentence --- rules 2,

→ Step 2

Subject (Pengkompil) Predicate [menukar bahasa paras tinggi kepada bahasa mesin].
Predicate: Verb (menukar) Object (bahasa paras tinggi) Conjunction (kepada) Adverb (bahasa mesin)
Adverb: Object (bahasa mesin)

→ Step 3

Menukar (bahasa paras tinggi, bahasa mesin)

b. Second, let use sentence no (4)
"Penganalisis Leksikal, Penganalisis Sintaks dan Penjana Kod Pertengahan merupakan fasa yang berlainan dalam pengkompil".

→ Step 1

Compound Sentence --- rules 24,

→ Step 2

Subject (Penganalisis Leksikal, Penganalisis Sintaks dan Penjana Kod Pertengahan)
Predicate [merupakan fasa yang berlainan dalam pengkompil].
Predicate: Verb (merupakan) Object (fasa yang berlainan) Conjunction (dalam) Adverb (pengkompil)
Adverb: Object (pengkompil)

→ Step 3

merupakan (penganalisis leksikal, fasa yang berlainan)
merupakan (penganalisis sintaks, fasa yang berlainan)

merupakan (penjana kod pertengahan, fasa yang berlainan)

7.0 SYSTEM ARCHITECTURE

The system consists of five modules to identify the five pola. The words of each segment are kept in a database. Some of the words may have more than one location. For example, the word, “telah”, is in the Adjunct, and in Conjunction. The architecture of the system is shown in Fig. 5. The sentences will be the input, and the segments with the properties of the Subject, Post-Subject, and the Predicate will be the output of the system.

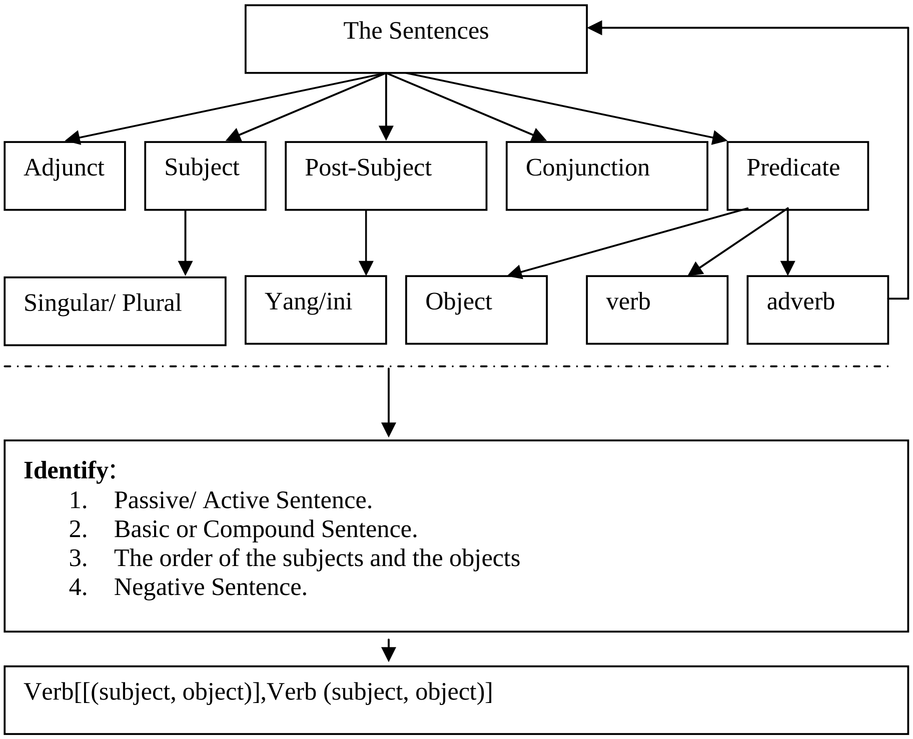
The Sentences
Adjunct
Subject
Post-Subject
Conjunction
Predicate
Singular/ Plural
Yang/ini
Object
verb
adverb
Identify:
1. Passive/ Active Sentence.
2. Basic or Compound Sentence.
3. The order of the subjects and the objects
4. Negative Sentence.
Verb[[(subject, object)],Verb (subject, object)]

Fig. 5: The system architecture based on pola sentence

7.1 Automata

The model is a finite automaton, mathematical model of a system represented as:

(Q, Σ, S, R), where
Q is a finite set of states
Σ is a finite set of input
S is the initial state
R is the transition relational which maps the input and states

There will be three levels for the processing. The first levels of the states are:
S1 = Adjuncts
S2 = Subject
S3 = PostSubject
S4 = Conjunction$_1$
S5 = Predicate
The transition diagram for the finite automata for the first level is,

→ (S1) → (S2)→ (S3) → (S4) → (S5)

The second level of processing is the Predicate (S5) processing. The states are:
S6 = Verb

S7 = conjunction
S8 = Object/ Agent
S9 = Adverb
The transition diagram is,

→ (S6) → (S7)→ (S8) → (S9)

The third level of processing is the Adverb (S9) processing and the states are:
S10 = Conjunction
S11 = Object
S12 = Conjunction
S13 = Adverb
The transition diagram is,

→ (S10) → (S11)→ (S12) → (S13)

8.0 TESTING AND RESULT

The pola grammar technique was developed to investigate the relationship between parsing and corpus method in NLP. Sagae et al [16] in similar experiments used 505 words and 118 sentences as the test set. The grammatical relations (GR) that were identified are subject, object, adjuncts and predicate nominal.

Yeh [17] divided the Grammatical Relations (GR) into the following sub-types:

1. Simple arguments: subject, object, indirect object, copula subject and object, expletive subject (e.g., “It” in “It rained today.”).
2. Modifiers: time, location and other modifiers.
3. Not simple arguments: arguments that syntactically resemble modifiers. These are location, objects, and also subjects, objects and indirect objects that are attached via a preposition.

Yeh [17] tested 1151 words on two systems, the transformation-based error-driven learning (TR) and memory-based learning system (MB). The result show that for simple arguments, the TR system performs better than the MB. The f-score for TR is more than 80% while the score for MB is 64%.

This paper discusses the experiments that used 19 abstracts thesis consisting 3604 words and 173 sentences to test the algorithm. The experiments objectives are:

i. To clarify the subject and Predicate,
ii. To analyze the Predicate, to identify the verb and object.
iii. To analyzed the Adverb, to identify more verb and object used in the sentence.

In this experiment, we compare the pola grammar technique (PG) with the parsing system for Malay language (PSM). Briefly, the PSM works as follow: the part of speech of the input text will be identified and then PSM parses the language based on the Malay language CFG. It consists a lot of ambiguities where there is no probabilistic model for the tree structure.

8.1 Result

Because precision and recall are often used as a parser evaluation metrics, it is common to envision a description of the syntactic constituent structure of sentences as the output [16]. Caroll et al [19] propose the precision and recall of grammatical relations be used for parser evaluation, and describe some advantages of using grammatical relations over other evaluation.

The result of the experiments is the number of instances of each grammatical relation (GR) that occurs in the test sentences. They are shown in Table 4 and 5. In order to compare the result done manually and the result produced by the algorithm, the standard information retrieval measures recall and precision is applied to measure how similar one such list is to another. The results are in Table 6, 7, 8 and 9. To perform comparisons between different variants of the grouping system, corresponding to the use of different combinations of linguistic modules, the F-score, which produces a number between precision and recall that is larger when the two measures are close together, is used.

Table 4: Number of instances of each GR in the test set

Grammatical Relation	Number of instances in test set
Adjunct	45
Subject	170
PostSubject	33
Conjunction	79
Predicate	170

Table 5: Number of instances of each GR in the test set (predicate)

Grammatical Relation	Number of instances in test set
Verb	121
Conjunction	74
Object	171
Conjunction	92
Adverb	167

Table 6: Results I for PG

Grammatical Relation	Recall	Precision	F-score
Adjunct	0.89	0.87	0.88
Subject	0.87	0.89	0.88
PostSubject	0.73	0.8	0.76
Conjunction	0.81	0.83	0.82
Predicate	0.86	0.86	0.86

Table 7: Results II for PG

Grammatical Relation	Recall	Precision	F-score
Verb	0.93	0.93	0.93
Conjunction	0.87	0.88	0.87
Object	0.79	0.78	0.78
Conjunction	0.72	0.74	0.73
Adverb	0.83	0.83	0.83

Table 8: Results I for PSM

Grammatical Relation	Recall	Precision	F-score
Adjunct	0.75	0.68	0.77
Subject	0.70	0.56	0.63
PostSubject	0.77	0.75	0.76
Conjunction	0.91	0.67	0.79
Predicate	0.81	0.76	0.79

Table 9: Results II for PSM

Grammatical Relation	Recall	Precision	F-score
Verb	0.7	0.62	0.66
Conjunction	0.67	0.66	0.665
Object	0.68	0.63	0.655
Conjunction	0.8	0.79	0.795
Adverb	0.82	0.8	0.81

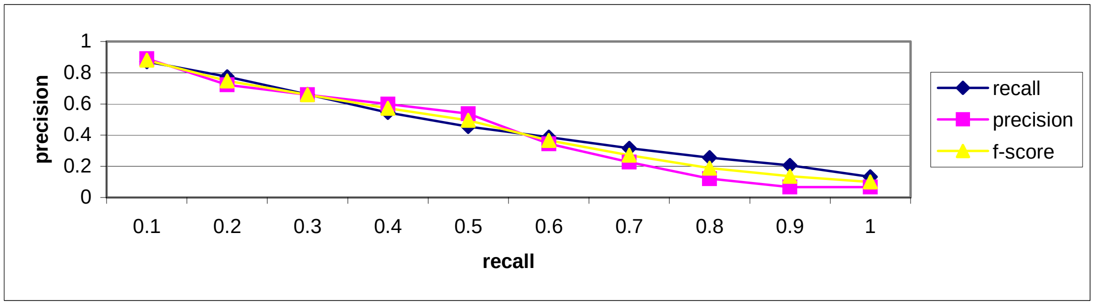

Fig.6: The graph for the recall and precision of the Subjects for PG

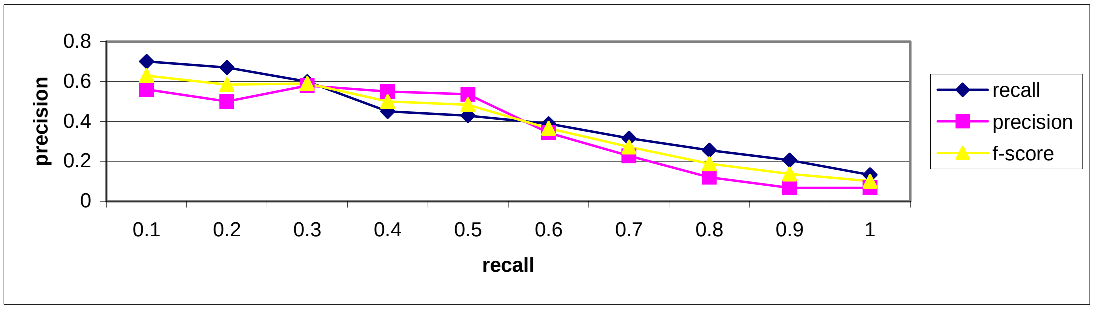

Fig. 7: The graph for the recall and precision of the Subjects for PSM

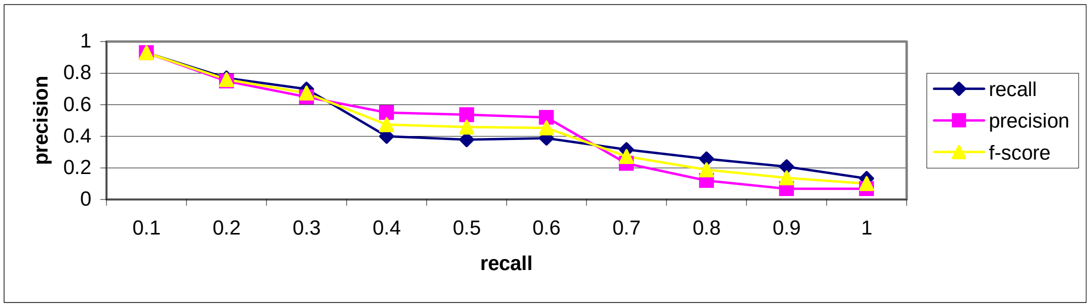

Fig. 8: The graph for the recall and precision of the verbs for PG

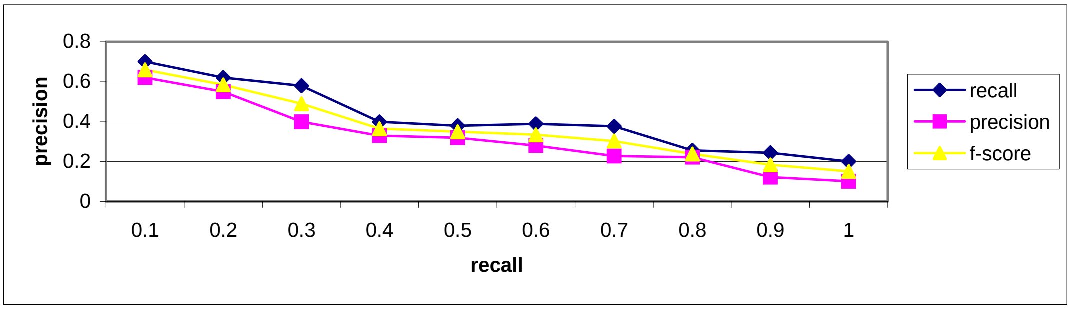

Fig. 9: The graph for the recall and precision of the verbs for PSM

9.0 DISCUSSION

The range of 73% to 93% of the F-score values, as shown in Fig. 6, 7, 8, and 9 show that the pola grammar techniques can be used to clarify the grammatical relations in Malay sentences. The relations are the functional structure of a language where it consists of a lexical mapping to tell the semantic role that the subject has. It extends the parsing system that relies very much on the part-of-speech of the tokens.

The unrecognized grammatical relations occurs, are the problem that were caused by the noun recognition, that do not has a fixed format and the post-subject problem which do not has a specific symbol to stop the pola. The summary of causes of the incorrect output is as follows.

i. The existing of the conjunction ‘dengan’ in the subject. The words that follow this word can either be as a postSubject or a subject.
ii. The nouns are varied and do not have a common pattern.
iii. The words ‘tidak’, to show a negative sentence do not locate in the right position.
iv. There is no specific “stop sign” for the postSubject except “,”, “ini”, and “itu”.
v. The verbs that act as a noun.

The examples of the problems are :

i. “Aliran kerja boleh ditakrifkan sebagai satu kaedah untuk mengautomasikan dan mengawal pergerakan proses”.
The word “Aliran kerja” is the subject where it is the title of the thesis. The program assumed that “aliran” is the subject and “kerja” is the verb.

ii. “Di samping itu, perlaksanaan sistem ini sedapat yang mungkin akan memudahkan perjalanan operasi”.
The first sentence consists of word “sedapat” after the subject and it should go to the postSubject, but it remains in the subject.

iii. “Data tahunan (unit masa besar) sebagai bahagian pertama dan data bulanan (unit masa kecil) sebagai bahagian kedua”.
The program processes the sentences inside a parathesis and produces wrong information. It identifies the subject as “Data tahunan (unit”. The rest of the sentence is the predicate.

REFERENCES

[1] A. Hassan, *Linguistik Am untuk Guru Bahasa Malaysia*, Fajar Bakti, Kuala Lumpur, 1980.

[2] A. M. Simin, *Discourse-Syntax of “YANG” In Malay (Bahasa Malaysia),* Dewan Bahasa dan Pustaka, Kuala Lumpur, 1988.

[3] A. Hj. Omar, *Morfologi-sintaksis Bahasa Malayu (Malay) dan Bahasa Indonesia: Satu Perbandingan Pola*, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1968.

[4] Lewis, M.Blanche, *Sentence Analysis In Modern Malay*, Cambridge, At The University Press, London, 1969.

[5] N. Othman, A Comparative Analysis of Malay and English Contrastive Discourse. PhD Dissertation, Boston University, 2000.

[6] S. Nathesan, Keberkesanan Penggunaan Kata Penanda Wacana dalam Penulisan, Pelita Bahasa, 1995, Sep 15-16.

[7] A. Hassan, *Tatabahasa Bahasa Melayu. Morfologi dan Sintaksis untuk guru dan pelajar*. PTS Publications & Dustributor Sdn. Bhd, Bentong, Pahang, 2002.

[8] N. S. Karim, *The Major Syntactic Structures of Bahasa Malaysia and their Implication of the Standardization of the Language*. PhD dissertation, Ohio University, 1975.

[9] N. S. Karim, Farid M. Onn, Hasihm Hj. Musa, A. H. Mahmood, *Tatabahasa Dewan*, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1993.

[10] N. S. Karim, F. M. Onn, H. Hj. Musa, A. H. Mahmood, *Tatabahasa Dewan Edisi Baru*, Dewan Bahasa dan Pustaka, Kuala Lumpur, 2004.

[11] M. Collins, *Head Driven Statistical Models For Natural Language Parsing*. Computer and Information Science. University of Pennsylvania, Pennsylvania, 1999.

[12] J. Carroll and T. Briscoe, "High Precision Extraction of Grammar Relations", in *Proceedings of 19th International Conference on Computational Linguistics (COLING)*. Taipei, Taiwan, 2002.

[13] A. Hj. Omar, *Nahu Melayu Mutakhir*, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1980.

[14] A. Hj Omar and R. Subbiah, *An Introduction To Malay Grammar*, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1995.

[15] H. Hj. Musa, *Sintaksis Bahasa Melayu*, Agensi Penerbitan, Kuala Lumpur, 1990.

[16] K. Sage, A. Lavie, and B. MacWhinney, "Combining Rule-Based And Data Driven Techniques For Grammatical Relation Extraction in Spoken Language", in *Proceedings of the Eight International Workshop on Parsing Technologies*, Nancy, France, 2003.

[17] A. Yeh, "Comparing two trainable grammatical relations finders", in Proceedings of the 18th International Conference on Computational Lingusitics (COLING 2000), Germany, 2000, pp. 1146-1150.

[18] C. K. Yeoh, Interaction of Rules in Bahasa Malaysia. PhD dissertation, University of Illinois at Urbana-Champaign, 1979.

[19] J. Caroll, Briscoe, and Sanfillipo. "Parser Evaluation : A survey and a new proposal". in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, 1998, pp. 447-454.

BIOGRAPHY

The first author is a lecturer in Computer Science Dept, Faculty of Information Technology and Information Science, Universiti Kebangsaan Malaysia, since 1999. Currently the author is pursuing his PhD program at Faculty of Computer Science and Information Technology, University Putra Malaysia. To date, the author has published 3 journal papers and more than 15 papers in proceedings related to documents processing, Automatic Marking and language processing. Fatimah Dato' Ahmad, Abdul Azim Abdul Ghani and Ramlan Mahmod are Associate Professor at Faculty of Computer Science and Information Technology, University Putra Malaysia. To date, most of their projects involve Natural Language Processing and Information Retrieval.