TRANSFORMING NOUN PHRASE STRUCTURE FORM INTO RULES TO DETECT COMPOUND NOUNS IN MALAY SENTENCES

Suhaimi Abdul Rahman[1] *and Nazlia Omar*[2]

[1]*Software Engineering Department, College of Information Technology*
Universiti Tenaga Nasional, Jalan IKRAM-UNITEN
43000 Kajang, Selangor, Malaysia

[2]*School of Computer Science,*
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Corresponding author: smie@uniten.edu.my[1]

ABSTRACT

This paper addresses the process of transforming the noun phrase structure form into a list of rules to detect compound noun words in Malay sentences. Rules are collection of word syntax that are derived from a specific resource (as defined in our study). Comprehension of the concept rule used in a system is important (i.e. using rules to find a list of compound nouns that may exist in a sentence). The noun phrase frame structure is a form that contains a list of noun modifier categories. The list of noun modifier categories is then divided into several sub-categories such as numeral, numeral classifier, appellation, etc. All categories are arranged in sequence based on correct grammar. The noun phrase frame structure is then used to analyse the sentence. The words in the sentence will be arranged according to their suitable noun modifier category as defined by the noun phrase frame structure. In terms of data requirements, we will only focus on examples of sentences that combine two noun phrases.

Keywords: Noun phrase structure form, rules, compound noun, noun modifier category, parts of speech, tokenizer.

INTRODUCTION

Recent research on Natural Language Processing (NLP) has been constantly growing, where one research area in text processing has been identified as a potential research which can be used to manage open text particularly in detecting a set of compound nouns retrieved from a sentence. According to English Club 15 Years (1997-2012), a compound noun is a noun made up of two or more words. In English, a compound noun always follows the syntax order [noun + noun] or [adjective + noun]. However, other combinations of compound nouns are available, such as [verb(-ing) + noun], [noun + verb(-ing)], [verb + preposition], [noun + prepositional phrase], [preposition + noun], and [noun + adjective]. By this definition, each compound noun acts as a single unit that can be modified by other nouns, adjectives, verbs, and prepositions.

The detection of compound nouns in a sentence is useful for the development of NLP application systems such as the detection of heads and modifiers in sentences, language translation systems, text summarization, word categorization, etc. This research work only focuses on the detection of compound nouns in Malay sentences. These sentences comprise a combination of two noun phrases also known as 'subject and predicate'.

It is significantly important to identify the type of data collection and preparation used. For this study, all the sentences and words were taken from a number of related sources that include magazines, newspapers, dictionaries, textbooks, and children's story books. The sentences and words need to be analysed and compiled before being entered into a database. During data manipulation, we developed a system to transfer all the words from a text file~~s~~ into a database. Meanwhile, another system was developed to assist in accelerating the process of data preparation and compilation.

RELATED WORK

As discussed by Hassan, (2004); Hassan (1992), Karim, Onn, Musa and Mahmood (2010) in the Malay language, compound nouns are based on a combination of words from any of the following categories: i) noun and noun, ii) noun and noun modifier, or iii) noun and non-noun modifier. The noun and noun category does not contain any modifier element~~s~~. Therefore, both words combined can lead to two different meanings i.e."*sama erti*"*(*synonyms) or "*lawan erti*" *(*antonyms). The noun and noun modifier category contains thirteen sub-categories defined as generation categories, instrument categories, position categories, body part categories, etc. In this category, the first noun can be modified by other nouns to form pairs of words in a compound noun. The noun and non-noun modifier category has six different types of sub-categories, namely: determiner, verb, adjective, adverb, preposition, and ordinal number. In this category, the first noun can be modified by other verbs, determiners, adjectives, adverbs, prepositions, and ordinal numbers, to form pairs of words in a compound noun. To construct a noun phrase structure form, we referred to several studies (Hassan, 2004; Hassan, 1992; Karim, Onn, Musa, & Mahmood, 2010; Guan, 2009; Taharin, Ja'afar, & Shukor, 2010; Chomsky, & Halle, 1991; Omar, 2009; Dinh Dien, 2002). Karim, Onn, Musa and Mahmood (2010) created a table consisting of a form to assign words. This form will assist in over viewing the structural pattern of Malay sentences added into the form. The word arrangement within the table is based on the three categories discussed earlier. However, the newly formed noun phrase structure still requires validation from linguists, especially in terms of checking for vague words with regard~~s~~ to placement words within the form and an identification of a group of words for a compound noun.

In another study, Dinh Dien (2002) used syntactic parsing of Vietnamese noun compounds to determine the noun compound of words. In terms of syntactic aspects, he defined that the first word of a noun compound 'must' be a noun, and the second word 'could' be a noun, verb, adjective, pronoun, preposition, number, etc. He used three different types of frame structures as follows: i) Frame structure of noun object, ii) Frame structure of verb object and iii) Frame structure of an adjective object. He used compound noun as a part of the process to detect the head and modifier of the words with the help of semantic relations between objects.

PARTS OF SPEECH IN MALAY SENTENCES

Part(s)-of-Speech (POS) is a linguistic category of words that is assigned to each word in a sentence. To obtain the POS for words in Malay, we used a dictionary compiled from Othman and Karim (2006). Based on this dictionary, all words were rewritten, together with their POS and noun modifier category. THE following are several examples of POS definitions using English words, (Chomsky, & Halle, 1991). The same definitions were used to describe the meaning of word categories in Malay.

i) **Noun:** A Noun is a word used to name a person, animal, place, thing, and abstract idea. (e.g., John, cat, box, desert, liberty, golf, etc.).

ii) **Verb:** A Verb is a word used to describe a noun's movement (action) or being (existence) (e.g., went, purred, is, etc.).

iii) **Pronoun:** A Pronoun is a word that can replace a noun or another pronoun (e.g. he, which, none, you, etc.).

iv) **Adjective:** An Adjective modifies a noun or a pronoun by describing, identifying, or quantifying words. An adjective usually precedes the noun or pronoun that it modifies (e.g. big, good, full, etc.).

v) **Adverb:** An adverb can modify a verb, an adjective, another adverb, a phrase, or a clause. An adverb indicates manner, time, place, cause, or degree, and answers questions (e.g. how, when, where, how much, quickly, loudly, here, etc.).

vi) **Conjunction:** A Conjunction is used to link words, phrases, and clauses (e.g. and, or, but, etc.).

vii) **Preposition:** A Preposition links nouns, pronouns, and phrases to other words in a sentence (e.g. at, under, over, of, to, in, out, beneath, beyond, for, etc.).

In order to label Malay words, POS, Table 1 shows several Malay POS with their corresponding English POS. We will use the Malay POS labels shown in Table 1 for the examples and discussions in our study.

Table 1

POS in Malay

POS in English	Label	POS in Malay	Label
Noun	N	*Kata Nama*	*KN*
Pronoun	Pro	*Kata Ganti Nama*	*KGN*
Verb	V	*Kata Kerja*	*KK*
Adjective	Adj	*Kata Adjektif*	*KAdj*
Adverb	Adv	*Kata Adverba*	*KAdv*
Conjunction	C	*Kata Hubung*	*KH*
Preposition	P	*Kata Sendi Nama*	*KSN*
Determiner	D	*Kata Penentu*	*KP*

NOUN PHRASE FRAME STRUCTURE

The noun phrase frame structure is designed to determine possible combinations of Malay words to produce Malay sentences. This form is used to find the compound noun words within a Malay sentence. A noun phrase frame structure contains two elements, namely noun modifier category and compound noun group. Analysing words using this noun phrase frame structure is useful to formulate rules which can be used in detecting compound nouns that may exist in a sentence.

a) Construction of a noun phrase

As discussed by Hassan (2004), Hassan (1992), Karim, Onn, Musa and Mahmood (2010), a noun phrase is made up of one or more words. The words must be in the correct order according to the grammatical structure of the target language being studied. In order to create a Malay noun phrase, it must have the following six elements (although not all elements must exist in the construction of a noun phrase):

i) *Kata bilangan* (Numeral)
ii) *Kata penjodoh bilangan* (Numeral Classifier)
iii) *Kata gelaran* (Appellation)
iv) *Kata inti* (Head)
v) *Kata penerang* (Modifier)
vi) *Kata penentu* (Determiner)

The combinations of word order are shown in Fig.1.

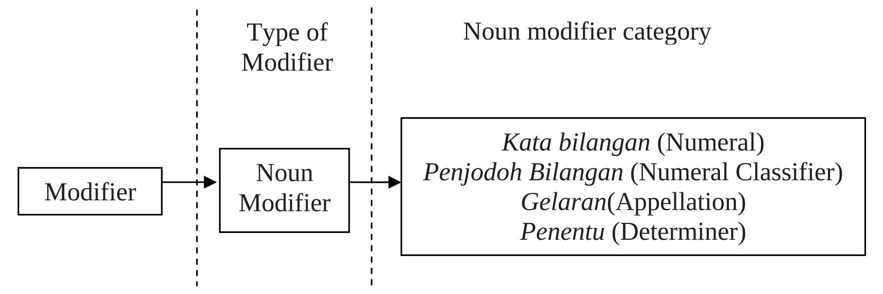


Figure1. Type of noun modifier.

Fig.1 shows the six types of noun categories that can be used to construct a noun phrase. In order to represent all six categories, we may arrange the words as follows:

Table 2

Noun Phrase Frame Structure

Noun Phrase							
Noun modifier category			Compound noun group			Noun modifier category	
Numeral	Numeral Classifier	Appellation	Noun	Noun	Modifier	Modifier	Determiner

Table 2 shows that words in the noun modifier category must come before the words in the compound noun group. The determiner should be positioned at the end of the sentence. This noun phrase frame structure serves as a guide~~s~~ to place words in the correct position; therefore, we know how this arrangement of words can be treated as a rule. To discuss how this noun phrase frame structure can be used to construct more noun phrases, some examples of Malay sentences are provided below:

Malay sentence 1:
"Rumah teres itu dibina sejak dua tahun lepas."
(The terrace house was built two years ago.)
Malay sentence 2:
"*Encik Ahmad seorang guru matematik di sekolah saya.*"
(Mr Ahmad is a math teacher at my school.)
Malay sentence 3:
"*Dua orang pelajar sedang tidur di atas meja tulis.*"
(Two students are sleeping on their desks.)

Based on the examples above, we can create a noun phrase frame structure to organise the words in their right positions. The examples of using noun phrase frame structure for positioning words are shown in Table 3. This frame structure was also discussed by some researchers (Hassan, 2004; Hassan,1992; Karim, Onn, Musa, & Mahmood, 2010).

Table 3

Noun Phrase Frame Structure for Malay Sentences 1, 2, and 3

Noun Modifier Category			Compound Noun			Modifier	Noun Modifier Category
Numeral	Numeral Classifier	Appellation	Noun	Noun	Modifier	Modifier	Determiner
-	-	-	*rumah*	-	*teres*	-	*itu*
dua	-	-	*tahun*		*lepas*	-	-
-	-	*Encik*	*Ahmad*	-	-	-	
seorang	-	-	*guru*	-	*matematik*	*di sekolah saya*	-
beberapa	*orang*	-	*pelajar*	-		*sedang tidur di atas meja tulis*	-
-	-	-	*meja*	-	*tulis*	-	-

b) **Construction rules**

Table 3 shows that word arrangements are based on the position of the noun modifier category. If the input word category does not match the type of noun modifier category defined in this noun phrase frame structure, the process will continue to search using the next noun modifier category. If the noun modifier category matches the input word category, the input word will be temporarily stored in this position. This process will stop at the determiner category. A compilation of the noun phrase structure from Table 3 is summarised below:

Noun phrase structure 1: Noun + Modifier + Determiner
Noun phrase structure 2: Numeral + Noun + Modifier
Noun phrase structure 3: Appellation + Noun
Noun phrase structure 4: Numeral + Noun + Modifier + Modifier
Noun phrase structure 5: Noun + Modifier

However, the list of noun phrases structures will be determined by observing examples from other Malay sentences. The more the examples observed, the higher the possibility of obtaining a new and improvised noun phrase structure. The concept of rules are also discussed by Michael (2002). To construct rules of a noun phrase structure, we will discuss the following:

Sentence 1:

Rumah teres itu dibina sejak dua tahun lepas." (The terrace house was built two years ago.)

From Sentence 1, we split the words (together with their own POS and noun modifier category). The POS and noun modifier category for each word will be obtained from a database. Table 4 shows the POS and noun modifier category for the sentence *"Rumah teres itu dibina sejak dua tahun lepas."*

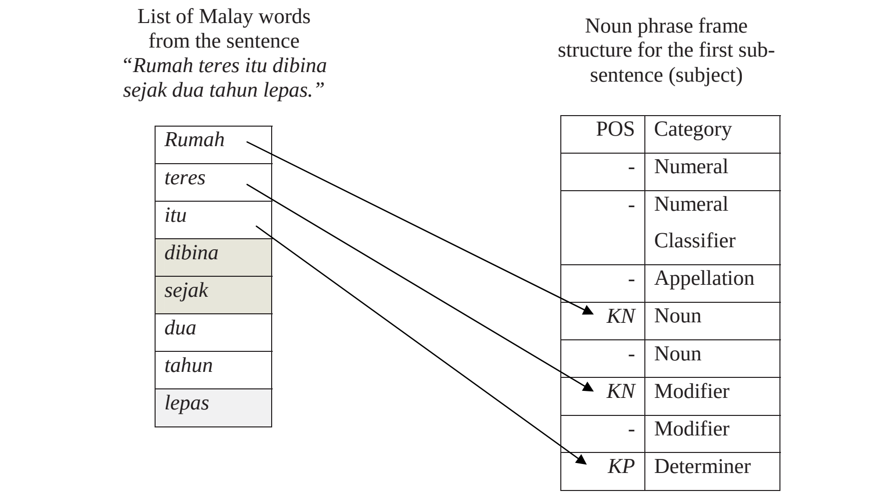
Table 4

POS and Noun Modifier Category for Sentence

Word	POS	Category
Rumah	*KN*	Noun
teres	*KN*	Modifier
itu	*KP*	Determiner
dibina	*KK*	-
sejak	*KH*	-
dua	*KN*	Numeral
tahun	*KN*	Noun
lepas	*KK*	-

The noun modifier category for this sentence is only occupied with the word "*dua*" (two) and the word *"tahun"* (year); this adheres with the numeral and noun type categories. The remaining words do not comply with the information obtained from a noun modifier category.

To formulate rules using a noun phrase frame structure, we divided the sentence into two sub-sentences where the first sub-sentence is called 'subject', and the second sub-sentence is called 'predicate'. The subject and predicate are analysed as follows:



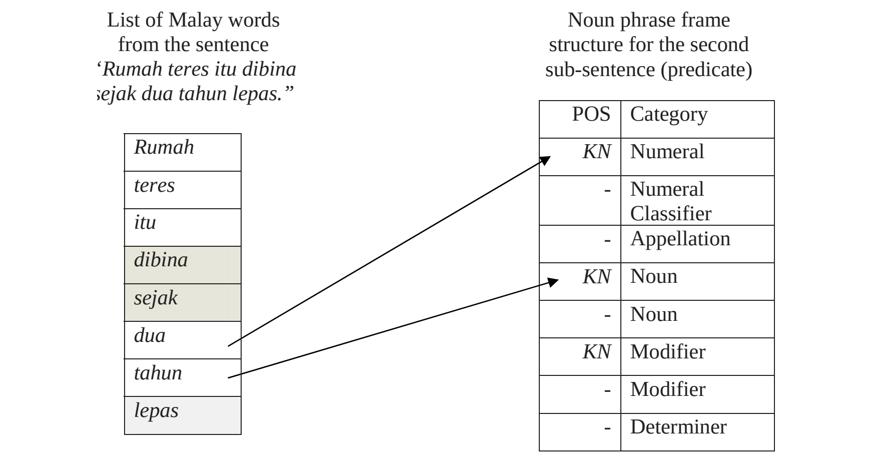
List of Malay words from the sentence *"Rumah teres itu dibina sejak dua tahun lepas."*

Rumah
teres
itu
dibina
sejak
dua
tahun
lepas

Noun phrase frame structure for the first sub-sentence (subject)

POS	Category
-	Numeral
-	Numeral Classifier
-	Appellation
KN	Noun
-	Noun
KN	Modifier
-	Modifier
KP	Determiner

Figure 2. Corresponding words with their POS and category.



List of Malay words from the sentence *'Rumah teres itu dibina ѕejak dua tahun lepas."*

Rumah
teres
itu
dibina
sejak
dua
tahun
lepas

Noun phrase frame structure for the second sub-sentence (predicate)

POS	Category
KN	Numeral
-	Numeral Classifier
-	Appellation
KN	Noun
-	Noun
KN	Modifier
-	Modifier
-	Determiner

Figure 2. (con't). Corresponding words with their POS and category.

In Figure 2, the sentence *"Rumah teres itu dibina sejak dua tahun lepas"* comprises two sub-sentences: subject and predicate. The sub-sentence "*Rumah teres itu"* will be named subject, while the remaining sub-sentence "*dibina sejak dua tahun lepas"* will be named predicate. Every word in the 'subject' will be checked against a database to find the POS and noun modifier category. If a word matches the POS and category in a noun phrase frame structure, the system will decide whether the sentence contains a compound noun. The same process is performed for the second sub-sentence; *"dibina sejak dua tahun lepas."* The sequence of words *"dibina [KK]", "sejak [KH]"* and *"lepas [KK]"*do not match the categories listed in the noun phrase frame structure. Therefore, we ignored these words and proceeded to the rest to examine the compound noun. The same process was used to check the other Malay sentences. However, in our research work, we focused on sentences with a combination of noun phrase and noun phrase, and not verb phrase, preposition phrase, and adjective phrase.

The following are some examples of rules constructed using a noun phrase frame structur as discussed above. The heuristic and relation rules explained in Michael (2002) will be used as a knowledge representation technique to construct the noun phrase structure rules defined in our study.

<u>Sentence 1:</u>

"Rumah teres itu dibina sejak dua tahun lepas." (The terrace house was built two years ago.)

The POS and noun modifier category for each word in this sentence is *"Rumah[KN] teres[KN] itu[KP] dibina[KK] sejak[KH] dua[KN][Numeral] tahun[KN] lepas[KK]."* Referring to the noun phrase frame structure depicted in Table 2, we can produce a list of rules that can be used to identify POS and noun modifier category in detecting a compound noun. Below are several examples of rules generated using this noun phrase frame structure:

<u>Rule 1:</u> KN + KN + KP[Detrminer].
(e.g., *Rumah[KN] + teres[KN] + itu[KP,* Determiner*])*

<u>Rule 2:</u> KN[Numeral] + KN[Numeral Classifier] + KN.
(e.g., *beberapa[KN,* Numeral*] + orang KN[*Numeral Classifier*] + pelajar[KN]*)

<u>Rule 3:</u> KN[Numeral] + KN[Numeral Classifier] + KN + KP[Determiner].
(e.g., *beberapa[KN,*Numeral*] + orang [KN,*Numeral Classifier*] + pelajar[KN] + itu [KP,*Determiner*]*)

Rule 4: KN[Appellation] + KN.
(e.g., *Encik[KN*, Appellation*]* + *Ahmad [KN]*)

Rule 5: KN[Numeral] + KN + KN.
(e.g., *Seorang[KN*,Numeral*]* + *guru[KN]* + *matematik [KN]*)

Rule 6: KN + KN
(e.g., *guru[KN]* + *matematik [KN]*)

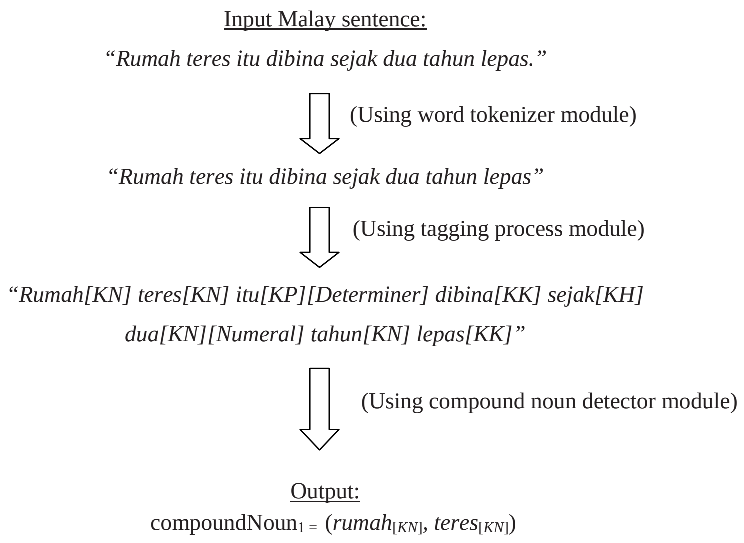
Rule 7: KN + KP[Determiner]
(e.g., *meja[KN]* +*itu [KP*, Determiner*]*)

All of the rules formulated from this analysis will be stored in a Malay noun phrase structure rule database for future reference. The more example sentences are analysed, the more rules and noun phrase frame structures can be constructed. The fundamental concept of compound nouns in Malay sentences was also discussed by Rahman, Omar and Che Hassan (2012); Rahman, Omar and A. Aziz (2011).

INPUT AND OUTPUT PROCESSES

Input and output processes importantly show the flow of data from one state to another. In order to produce the correct results, we need to identify a suitable data type that can be accepted by the system. The input data will proceed to the next process known as the compound noun detector. To obtain a compound noun from a sentence, we used the above-mentioned concepts. Below is an example of the input and output processes that will be used in our research.

Figure 3 shows that there are three main processes used to detect a compound noun, namely: word tokenizer module, tagging process module and compound noun detector module. The word tokenizer module chunks the whole sentence into its own words. This process also removes unnecessary symbols and punctuation marks from the input sentence. This is done for the purpose of computer data processing. The second process is the tagging module. This module labels suitable POS and noun modifier category for each word in the input sentence. The POS and noun modifier category is obtained from a database. Once the data is obtained from the database, the system will manage the words and generate an output based on the form defined in our research requirement. The third module is known as the compound noun detector. This module identifies whether possible compound nouns exist in a sentence. To achieve this goal, we performed a meticulous study to identify appropriate techniques to be used in our problem solving. This was performed by implementing a concept of noun phrase frame structure that could produce rules. These rules are beneficial when used to detect compound nouns.



Input Malay sentence:

"Rumah teres itu dibina sejak dua tahun lepas."

(Using word tokenizer module)

"Rumah teres itu dibina sejak dua tahun lepas"

(Using tagging process module)

"Rumah[KN] teres[KN] itu[KP][Determiner] dibina[KK] sejak[KH] dua[KN][Numeral] tahun[KN] lepas[KK]"

(Using compound noun detector module)

Output:

compoundNoun$_{1}$ = (*rumah*$_{[KN]}$, *teres*$_{[KN]}$)

Figure 3. Input and Output Processes.

CONCLUSION

Generally, this paper does not discuss any evaluation of test results to measure the accuracy of the results of compound nouns. The fundamental concept of noun phrase was discussed in which a detailed review of compound nouns in Malay sentences was performed.

We discussed the most important approach to recognize compound nouns in Malay sentences. In order to find a compound noun, we used a combination of three modules, namely: word tokenizer, tagging process and compound noun detector. Although each module has its own task, they can link with each other. An output from one process becomes the input for another. This process flow continues until a compound noun result is produced.

The rules, obtained from a noun phrase frame structure are vital in detecting compound nouns in Malay sentences. In order to obtain more rules, further analysis of other Malay sentence examples is required particularly in sentences that are derived from a combination of two noun phrases.

Research on compound nouns for Malay is important and should become a specific research area for Natural Language Processing (NLP). This research area mainly focuses on analysing and manipulating a collection of texts or lexicons to be used in developing a language application system.

ACKNOWLEDGEMENT

We would like to thank Ministry of Higher Education (MOHE) for the Fundamental Research Grant Scheme (FRGS, 2011).

REFERENCES

Abdullah Hassan. (2004). *Tatabahasa Bahasa Melayu* (edisi ke-4). Kuala Lumpur: PTS Publications & Distributors.

Abdullah Hassan. (1992). *Linguistik am* (Edisi ke-10). Kuala Lumpur: PTS Profesional Publishing.

Arbak Othman, & Nik Safiah Karim. (2006). *Kamus komprehensif Bahasa Melayu* (Cetakan kedua). Petaling Jaya: Penerbit Fajar Bakti.

Asmah Omar. (2009). *Nahu Melayu mutakhir* (Edisi ke-5). Kuala Lumpur: Dewan Bahasa dan Pustaka (DBP).

Chomsky, N., & Halle, M. (1991). *The sound pattern of English* (2nd ed.). Press, Cambridge, MA: MIT.

Dinh Dien. (2002). Cognitive linguistics approach to Vietnamese noun compounds. *Mon-Khmer Studies, 32,* 145-162.

English Club 15 Years. (1997-2012). *Compound nouns.* Retrieved from www.englishclub.com/ grammar / nouns-compound.htm

M. Taharin, R. Ja'afar, & N. A. Shukor. (2010). *Tesaurus Bahasa Melayu Dewan.* Kuala Lumpur: Dewan Bahasa dan Pustaka (DBP).

Michael, N. (2002). *Artificial Intelligence: A guide to Intelligent systems* (2nd ed.). Kuala Lumpur: Pearson Education.

Nik Safiah Karim, Farid M. Onn, Hashim Musa, & Abdul Hamid Mahmood. (2010). *Tatabahasa Dewan* (Edisi ke-3). Kuala Lumpur: Dewan Bahasa dan Pustaka (DBP).

O.C. Guan. (2009). *Kuasai struktur ayat Bahasa Melayu.* Kuala Lumpur: Dewan Bahasa dan Pusataka (DBP).

Suhaimi Ab Rahman, Nazlin Omar, & Noor Baizura Che Hassan. (2012). Construction of compound nouns (CNs) for noun phrase in Malay sentence. *Proceeding of CAMP 2012*, 22-25.

Suhaimi Ab Rahman, Nazlin Omar, & Mohd Juzaidi Ab Aziz. (2011). Transformation of Malay head modifier noun phrase into a thematic relation structure. *Proceeding of ICEEI 2011*,1775-1779.