VOL. 11, NO. 22, NOVEMBER 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

MALAY PARSE TREE SENTENCE VISUALISATION (BMTUTOR): COMPONENTS AND MODEL

Yusnita binti Muhamad Noor and Zulikha binti Jamaludin
School of Computing College of Arts and Sciences Universiti Utara Malaysia Sintok, Kedah, Malaysia
E-Mail: s92715@student.uum.edu.my

ABSTRACT

Language researchers introduce parse tree sentence visualisations to help users understand sentence structures. However, language research in parse tree sentence visualisation for Bahasa Melayu (Malay language) still has not attracted enough researchers to produce a model and a prototype which enable the visualisation of Bahasa Melayu (BM) sentences. Trends for BM language researches are mostly geared towards developing parsers (sentence checkers) for BM sentences. The learning of BM words has been achieved manually up to now. Sentence formation has to be learned at the school level. Thus, BMTutor has been introduced to help students in learning Malay sentences and word classes through a computerised visualisation method. Researchers have identified the difficulties faced by students in understanding word class and sentence structures. This will certainly benefit students especially those who have difficulties in understanding sentence structure. The BMTutor helps to understand the BM sentence structure by combining four components, which are 1) sentence checker; 2) sentence corrector, 3) parse tree sentence visualization, and 4) word attribute components. During the development phase, the result of the tested prototype showed 87.5% of the BM sentences were successfully parsed with only one parse tree visualisation output for each sentence used.

Keywords: BMTutor, parse tree sentence visualisation, sentence checker, sentence correction, word attribute components.

INTRODUCTION

In explaining a language structure, language researchers use a parse tree representation, which involves the grammar formation amongst the words. For Bahasa Melayu (BM), the same approach can be seen in Yusoff [1], Karim, Onn, Musa and Mahmood [2], Hassan, Jaya Rohani, Ayob and Osman [3]. Since no computer-based system has been introduced among others, the sentence representations were administered in a paper-based format. This format requires more space and effort. Most importantly, there is a lack of emphasis by the researchers in BM sentence processing as described in the computational linguistics and natural language articles [4, 5, 6]. Thus, we believe by introducing BMTutor, BM can be more progressive in computer-based language processing.

BMTutor can be used as a tool to help students learn BM sentences better since there are evidence on poor performance in the BM grammatical sentences amongst the students [7, 8]. This grammar issue is due to the difficulty in understanding the grammatical structure [9]. Daing Melebek [10] states that students have faced difficulties in understanding the BM sentence structure due to inability to use the correct word class which eventually affect their sentence construction skills.

This paper discusses the model of the BMTutor for BM parse tree sentence visualisation (PTV). The BMTutor aims to assist users, especially the students, to explore and learn about the structure of a sentence through phrase structure formation and word class visualisation. The objective is to improve the correct use of phrases and word classes in BM language. BMTutor can be used as a guide for the students in better understanding the BM language sentence structure. This is in line with the role of BM as the important language in Malaysia and it is compulsory for all Malaysians to use BM in their daily communication [9]. This has also been the motivating factor for selecting BM as the target language in this study. To date, the available computerised applications that focus on BM are still at its infancy compared to those of other languages, especially English [4, 5, and 6].

RELATED MODEL

The Structure-String Tree Correspondence (SSTC), semantic, and syntax parser models were analysed to get the components involved. Only these three models are related to the PTV and BMTutor research scope. The comparison amongst these three models were analysed as shown in Table-1.

Table-1. Model comparison.

Characteristic	SSTC [13]	Syntax parser [14]	Semantic parser [15]
Aim	For machine translation	For consolidation of language in cognitive science	To get an understanding of mathematics learning
Components	1. Sentence 2. Phrase 3. Database (example-based) 4. Parser	1. Word 2. Phrase 3. Characteristic structure 4. Grammar rules 5. Lexical rules	1. Parser 2. Lexicon 3. Grammar

The components needed for the development of a parser or a language processing application absolutely must contain the lexicon or database and representatives of the words and phrases that represent a sentence. Table-1 provides a summary of the components involved in processing a sentence that can be summarised as 1) word/phrase/sentence, 2) lexicon/database, 3) parser, and 4) grammar/lexical rules. Therefore, based on these four components, a model of the BMTutor was developed. The uses of these components have been combined in the BMTutor as a sentence checker and PTV. Besides that, a sentence corrector and word attribute components have also been added as discussed in the next section.

BMTUTOR COMPONENTS AND MODEL

The BMTutor is a PTV package combined with a sentence checker, sentence correction, PTV, word attribute components and PTV for sentence examples. The sentence checker needed to be included to produce a parse tree solely for a grammatical sentence. After the checking process, a sentence correction will be proposed for any incorrect sentence entered. The review and recommendation processes require rules according to the BM CFG to generate the PTV. There is a list of words, word classes and types of phrases saved in the BM CFG. The PTV displays the type of phrase, word class and word for the word input. Each word in the PTV can be selected to determine the combined set of components including a list of example sentences that has links to the formation of the new PTV as shown in Figure-1.

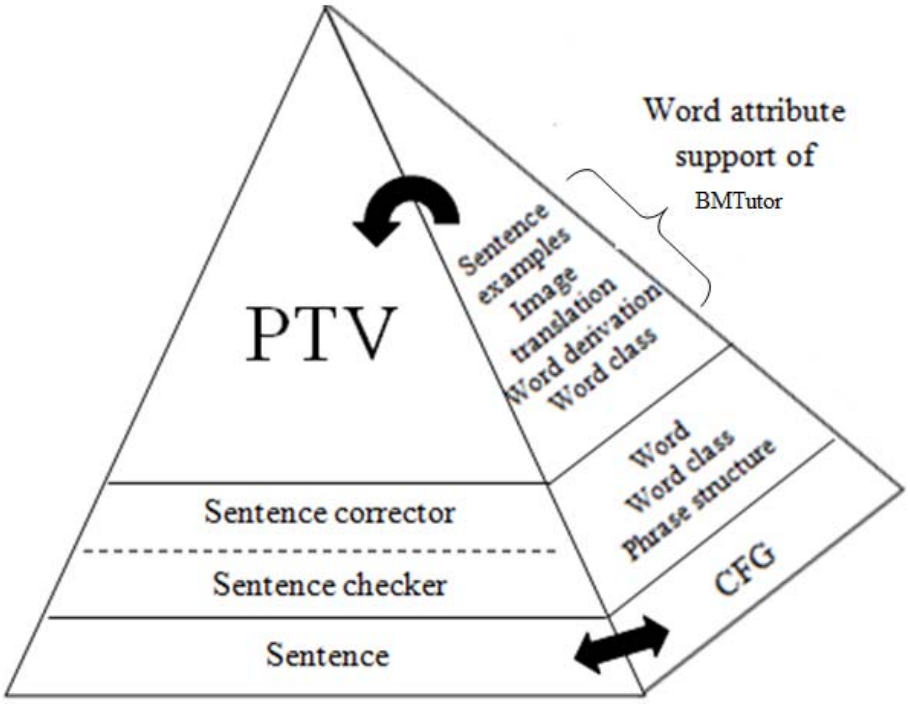

Figure-1. Pyramid model of the BMTutor.

The Pyramid model of BMTutor (Figure-1) shows that PTV needs sentence checker and sentence corrector to analysing the word, word class and phrase structure. Analysing process will invoved CFG as the

database. In addition, word attribute support are combined as a new PTV will be produced. Detail of functionalities of each components in Figure-1 are described in Table-2.

Table-2. Detail of functionalities of each BMTutor components.

Component	Detail of functionalities
1. Parse tree visualisation The parse tree is divided into subject and predicate in a hierarchical design. PTV is produced based on sentence input by the user and sentence examples produced by the system.	A. Parse tree for input sentence The parse tree will be visualised only when the entered sentence is correct according to the sentence structure rules for a declarative sentence consist of 14 words or less. The processes involved in producing the PTV are as follows: 1. Sentence checker will start by counting the number of words according to the rules provided. 2. Sentence condition will be checked for each word and its appropriate word class to ensure that the sentence received for further analysis is only for declarative sentence. 3. Each word will be matched with an appropriate word class. 4. Syntax checking is performed by matching each syntax in the input sentence with the appropriate rules provided. 5. For a successfully matched syntax, the respective PTV is generated and displayed. Otherwise, a sentence correction is proposed. 6. Each node in the PTV is ambedded with a hyperlink connected to the respective word attribute components. B. Parse tree from sentence examples A list of sentence examples is included in the word attribute components repository. The sentences are retrieved according to the word selected by the user in the existing PTV. Each sentence examples has a hyperlink that enable the generation of a new PTV.
2. Sentence checker	The sentence checker performs word class tagging and sentence rules matching.
3. Sentence correction	In the case of unmatched rules, a sentence correction will be proposed. The combination of rules used in the input sentence will be matched with the rules available in the database.
4. Word attribute components	As mentioned previously, each node in the PTV for the input sentence will have a hyperlink to the repository of the word attribute components. The attributes include word class, word derivation, translation, image and sentence examples. Each attribute is displayed according to the word choice made by the user in the PTV. Since each sentence example has hyperlink to make a new PTV, users may view different type of sentences in different contexts for each word selected.

Before finishing the program interface and encoding as well as completing the algorithms, this study proceeds to generate the parse tree text analysis. The overview of the applied programming process is shown in Figure-2.

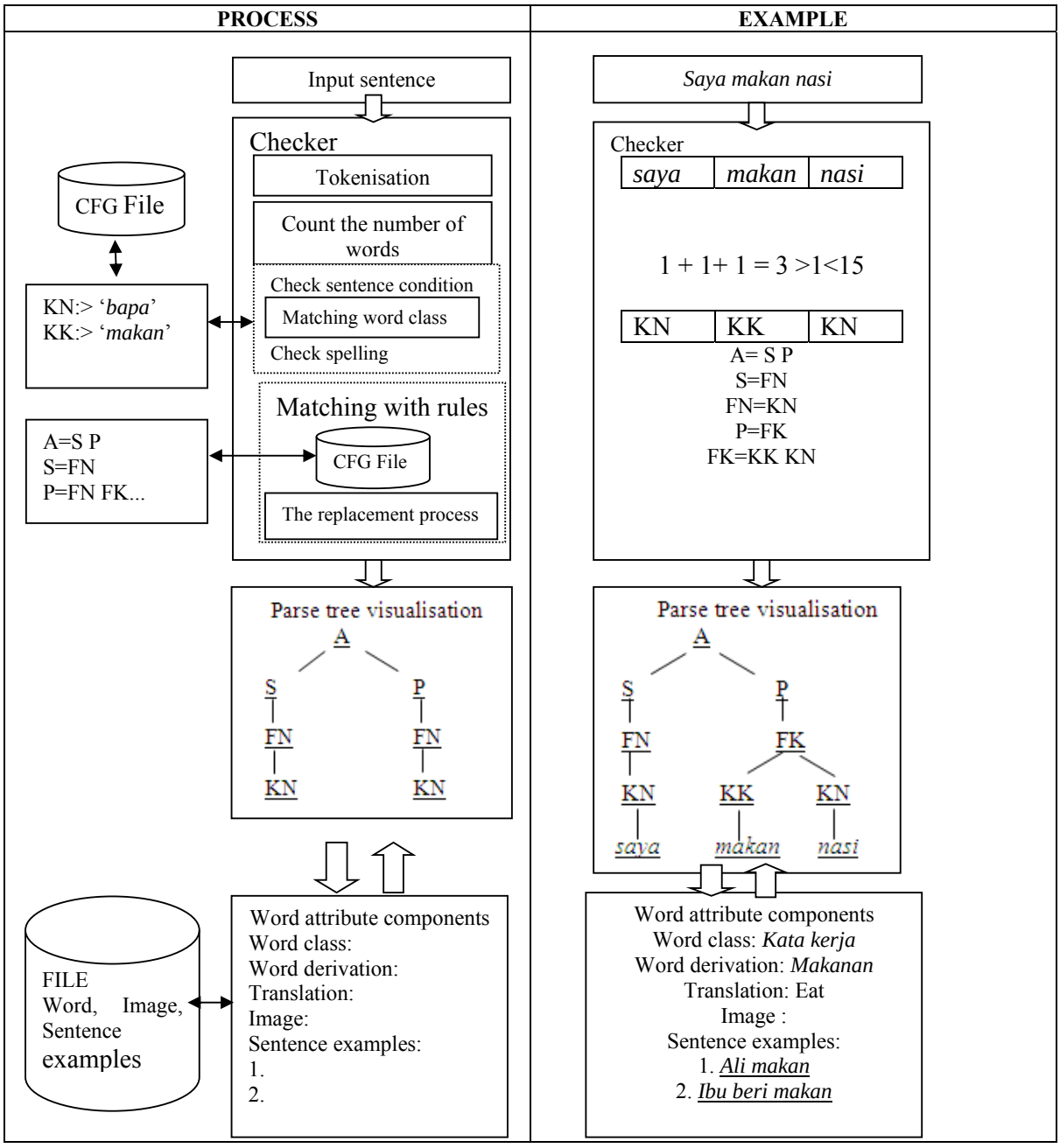

Figure-2. The process of analysing a sentence in the BMTutor.

The same process is also used to review and suggest sentence corrections. In the review process (Figure-2), the checking of spelling, sentence condition and matching with rules are part of the selection process. These processes are performed based on the conditions given in the source code within the scope of this study. The replacement process will play a role when the sentence during "matching with rules" does not meet the correct CFG. The replacement process is also needed for the sentence correction algorithms. Further visualisation processes will not be performed unless an error message and correct sentence suggestion are displayed to the user. During the "word attribute components" process, the process flow shows that there is a flow in and out of the processes because the output will contain sentence examples that have a link to create a new PTV.

The matching process with the repository also shows the flow of repetitions performed by the prototype because every word will be referred to, one by one, until all the words are matched.

The BMTutor interface is divided into two parts in one screen. The left part is to visualise the parse tree; whilst the right is set to display the word attribute components. The interface is shown in Figure-3.

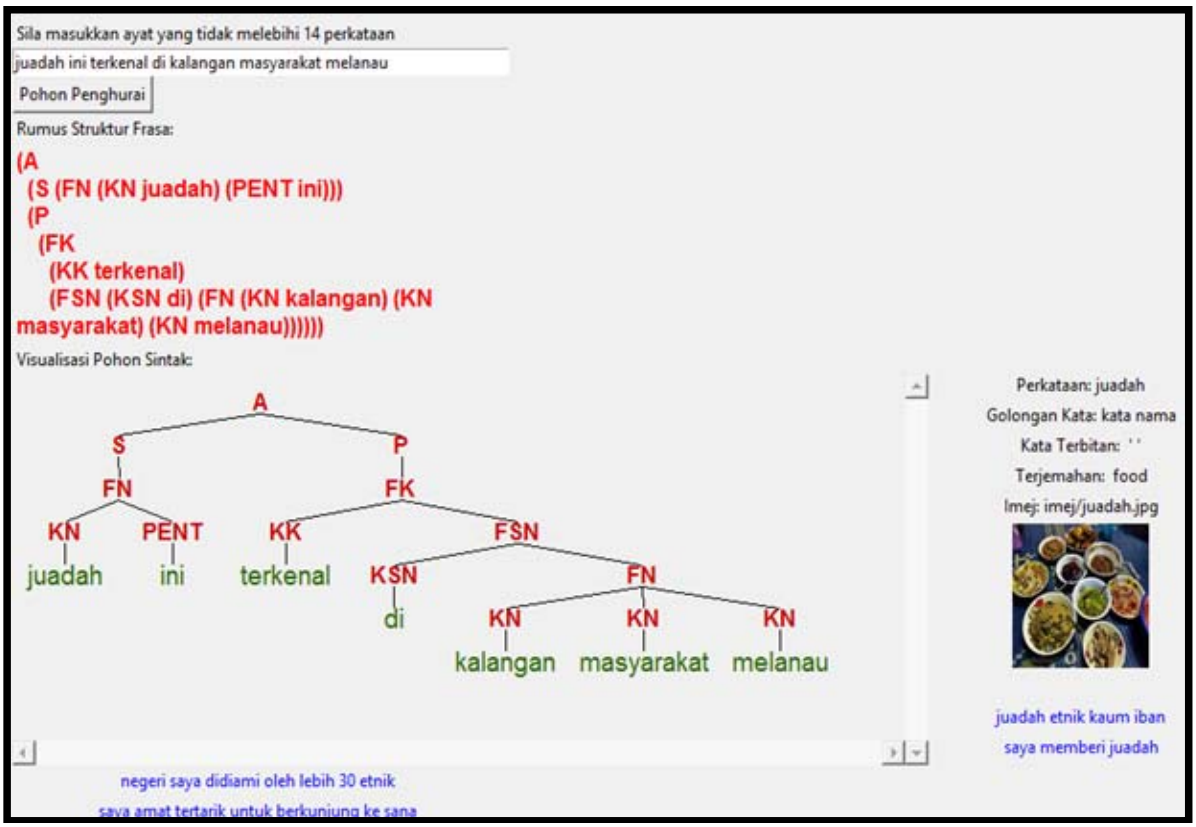

Figure-3. BMTutor interface.

PROTOTYPE TESTING

A total of 1884 declarative sentences with no more than 14 words were collected from the form 1 to form 5 BM textbooks. All sentences were divided into three categories to be used during the prototype development, training and prototype testing. Each category involved 628 sentences.

As the BMTutor prototype is still in the process of improvement, only the output from the development phase is shared in this paper; whilst the weaknesses will be improved in the next process. The outputs from the training and testing phases will be presented in the next study.

For the first time, to test the developed prototype, a total of 50 types of sentences were selected randomly for the purpose of collecting the related rules. The rules were obtained through the paper-based parse tree sketch of each sentence. From the collected sentences, 20 sentences were collected randomly to test the PTV and 15 invalid sentences to test the sentence correction. The results indicate that the prototype can produce a good output with correct sentence correction for all the 15 sentences; whilst for the PTV, the prototype successfully produced 13 correct PTVs with one output (65%) and the remaining 7 sentences (35%) produced multiple outputs.

For the second test, the CFG derived from the 628 sentences were grouped in the development category. 10% of these sentences were taken and tested because the number of samples to be measured or used is preferably 10% of the total sample [11, 12]. 10% of the 628 sentences were 63 sentences that were randomly selected from the collected sentences. The same method will be used in the prototype training and testing.

From the 63 tested sentences, 56 sentences were successfully produced with the PTV of only 1 output. The other 8 sentences were also successfully displayed. However, there was more than one output because of the ambiguity in the structure of the sentences. Each node in the PTV could be selected to view the word attribute components. From the list of sentence examples, the new PTV was successfully displayed when the link was selected. The displayed output is shown in Table-3.

Table-3. PTV output.

Output	PTV	Word attribute components	PTV for sentence examples
1 PTV	56	/	/
>1 PTV	8	/	/

The total sentences that were correctly parsed (B) were divided by the total sentences (A) as shown in Table-4. This indicates that the prototype had successfully visualised the sentence to produce only one PTV of 87.5%. This also shows that the rules used in analysing the sentences can be used by the prototype as well. The result was better when compared to the use of only 8 sentences or 12.5% where it showed that the prototype produced more than one result because of the ambiguity and other factors.

Table-4. Results from the development phase.

A	B	(B/A) *100%
64	56	87.5
64	8	12.5

The results of the performed experiments have shown an increase from 65% to 87.5% PTV accuracy with one output. This increment demonstrated that the prototype managed to successfully produce PTVs based on the rules collected from 628 sentences compared to only 20 sentence outputs which comprised the rules collected from only 50 sentences.

CONCLUSION AND FUTURE WORKS

Overall, this study has reported on an ongoing project regarding the development of the BMTutor. The BMTutor was developed to fulfill the needs of secondary school students in learning the BM sentence structure and to solve issues regarding their weaknesses in essay writing especially in understanding the correct use of phrase structure and combination of word classes. Learning through the formation of phrase structure, combination of word classes, combination of word components and sentence correction are provided in the BMTutor. Hence, a pyramid model of the BMTutor has been proposed. The components involved in the model are 1) sentence checker, 2) sentence corrector, 3) PTV, and 4) word attribute components. The model proposed has guided the development of the BMTutor and the evaluation showed that the prototype can produce a positive output with the accuracy of the output being more than 50%.

REFERENCES

[1] Z. Yusoff, Cintailah bahasa kita, suatu tanggapan linguistik berkomputer. Universiti Sains Malaysia: Pulau Pinang, Malaysia, 1998.

[2] N. S. Karim, F. M. Onn, H. Musa, & A. H. Mahmood, Tatabahasa Dewan Edisi Ketiga, Dewan Bahasa dan Pustaka: Kuala Lumpur, 2009.

[3] A. Hassan, S. L. Jaya Rohani, R. Ayob, & Z. Osman, Sintaksis, Siri pengajaran dan pembelajaran Bahasa Melayu. PTS Professional Publishing Sdn. Bhd.: Kuala Lumpur, Malaysia, 2006.

[4] M. J. Ab Aziz. 2007. Pengkomputeran Linguistik Bahasa Malaysia [Online]. Retrieved Dec 28, 2010, Available: http://www.ftsm.ukm.my/programming/prosiding-atur07/08-Juzaiddin.pdf.

[5] N. A. Bakar, M. Salaebing, S. Salleh, N. M. Rodzes, & F. M. Ishak. 2006. Penggunaan komputer dalam pengajaran bahasa. Retrieved Dec 28, 2010, available http://202.28.66.7/smuhammad/pdf/Penggunaan%20Komputer%20dlm%20pengajaran%20bahasa.pdf.

[6] S. Ramli, "Reka bentuk dan implementasi suatu penghurai bahasa Melayu menggunakan sistem logik selari", M.S. thesis, Universiti Putra Malaysia, Selangor, 2002.

[7] Z. Ahmad & N. H. Jalaluddin. Incorporating structural diversity in the Malay grammar. GEMA Online™ Journal of Language Studies, 12(1), Special Section, 17-34. 2012.

[8] N. H. Jalaluddin, J. Kasdan & Z. Ahmad. Sosiokognitif pelajar remaja terhadap Bahasa Melayu. GEMA Online™ Journal of Language Studies, 10(3), 67-87. 2010.

[9] A. C. Bagavathy. Mengatasi Kelemahan Murid Menguasai Aspek Tatabahasa Dalam Bahasa Melayu Melalui Cara Permainan Bahasa. Prosiding seminar penyelidikan pendidikan IPBA 2005, 50-58, 2005.

[10] A. R. Daing Melebek, "Perubahan struktur kata tunggal Bahasa Melayu mengikut aliran", PhD thesis. Universiti Putra Malaysia, 2004.

[11] M. Y. Ahmad, N. S. Karim, M. I. Takiah, J. Mansor, I. Juneidah, & Y. Abd. Kahar. Laporan kaji selidik penggunaan Bahasa Malaysia di institusi penggajian tinggi. DBP: Selangor, 1992.

[12] A. G. Mohd Najib, & S. Salehudin. 2007. Konsep Kendiri Pelajar: Kajian di Sekolah Menengah Sekitar Bandaraya Kuching, Sarawak. Retrieved January 24, 2011,available: http://www.ipbl.edu.my/bm/penyelidikan/seminarpapers/2007/Edpsychology/salehudinIPBLfp.pdf.

[13] M.H. Al-Adhaileh, & T.E. Kong. 1998. A flexible example-based parser based on the SSTC. Proceeding COLING '98 Proceedings of the 17th International conference on Computational linguistics, 1, 687-693. doi: 10.3115/980845.980960.

[14] Murugesan, A., & Cassimatis, N. 2006. A model of syntactic parsing based on domain-general cognitive mechanisms. Prosiding 8th annual conference of the Cognitive Science Society. Vancouver Canada.

[15] Peters, M. 2008. The Development of a Semantic Model for learning Mathematics. Proceedings of the British Society for Research into Learning Mathematics. Dicapai pada 15 Januari 2015, daripada http://www.bsrlm.org.uk/IPs/ip28-2/BSRLM-IP-28-2-14.pdf.