SIMPLE MALAY TO ENGLISH TRANSLATOR

By

SALMI SUZANA AHMAD FADZIL

Dissertation submitted in partial fulfillment of
the requirements for the
Bachelor of Technology (Hons)
(Information Communication Technology)

DECEMBER 2005

Universiti Teknologi Petronas
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

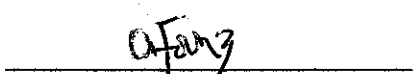
SIMPLE MALAY TO ENGLISH TRANSLATOR

by

Salmi Suzana Ahmad Fadzil

Dissertation submitted in partial fulfillment of
the requirements for the
Bachelor of Technology (Hons)
(Information Communication Technology)

Approved:



Amy Foong Oi Mean
Project Supervisor

UNIVERSITI TEKNOLOGI PETRONAS
TRONOH, PERAK

December 2005

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

Salmi Suzana Ahmad Fadzil

ABSTRACT

In the modern world, there is an increased need for language translation. Attempts of language translation are as old as computer themselves. Machine translation is the attempt to automate all, or part of the process of translating from one human language to another language. Machine Translation involves translating from a source natural language to a target language. Machine Translation is hard because structures in one human language often do not correspond in a simple way to structures in another. This paper represents a prototype of a Simple Malay to English Translator. This translator is developed to translate simple Malay sentence to English sentence since there is not many Malay-English translator available. The main tools that will be used for the project development are Java Language, Forte for Java 4.0 Community Edition and Microsoft Notepad version 5.1. From the research done, a dictionary that is used for a machine translator is usually being created in a notepad file for easy retrieval compared to using Microsoft Access of other database application. The ambiguity problem would not be addressed in this project. Hence, the goal of the project is to translate syntactically correct and the semantic factor is not taken into consideration.

ACKNOWLEDGEMENT

First and foremost, I would like to extend my deepest gratitude to my FYP supervisor, Ms. Amy Foong Oi Mean. The guidance that Ms. Amy has given me is priceless and valuable. As my supervisor, Ms. Amy has the patience and charisma in guiding me throughout the FYP. Thank you for your continuous support and guidance.

I would like to express my heartfelt thanks to the UTP lecturers that was involve in the questionnaire distribution. Thank you for your time and cooperation. I am also profoundly grateful to Ms. Norfaazah Yacob for the books that were lent to me as my reference. Without the books, I will not be able to do the analysis of the Malay language. My deepest gratitude goes to my friends who have helped me in giving ideas and assisting me when I face problems. Thank you for your help and support.

It would be impossible to complete the Final Year Project without the helps from all of them. Last but not least, to those who assist me directly or indirectly in making this project into a success. Thank you very much, may God repays for your kindness.

TABLE OF CONTENTS

CERTIFICATION.....i

ABSTRACT.....iii

ACKNOWLEDGEMENT.....iv

CHAPTER 1: INTRODUCTION

1.1 Background of Study.....1

1.2 Problem Statement.....2

1.3 Objectives and scope of study.....3

CHAPTER 2: LITERATURE REVIEW.....4

2.1 Machine Translation.....4

2.2 Related Work.....4

2.3 Translation Problem.....5

2.4 Source Language Analysis.....6

2.5 Translation Method.....7

2.6 Translation Approaches.....8

2.7 Design of the Bilingual Dictionary.....11

2.8 Issues In the Construction of the Bilingual Dictionary.....13

CHAPTER 3: METHODOLOGY/PROJECT WORK

3.1 METHODOLOGY.....15

3.1.1 Planning.....15

3.1.2 Analysis.....15

3.1.3 Design.....19

3.1.4 Development.....20

3.1.5 Testing/Delivered.....20

3.2 SIMPLE MALAY TO ENGLISH TRANSLATOR ARCHITECTURE.....21

3.3 MALAY – ENGLISH DICTIONARY DATABASE.....22

3.4 TOOLS.....23

CHAPTER 4: RESULTS AND DISCUSSION....24
4.1 Results and Discussion....24
4.2 Unit Testing....25
4.3 Integration Testing....26
4.4 Tree Diagram of Malay Sentence Structures That Can Be Translated....27
4.5 User's Feedback....33
4.6 System Prototype Translation Accuracy....33
4.7 Sample Snapshots of Output....35

CHAPTER 5: CONCLUSION AND FUTURE RECOMMENDATION....38
5.1 Conclusion....38
5.2 Recommendation....38
5.2.1 Dictionary....38
5.2.2 Simple Malay to English Translator....38

REFERENCES....39
APPENDICES....42
Appendix I: Project Timeline....43
Appendix II: Questionnaire....44
Appendix III : Abbreviation Explanation....50

LIST OF FIGURES

Figure 1: Pola 1.....17
Figure 2: Pola 2.....18
Figure 3: Pola 3.....18
Figure 4: Pola 4.....18
Figure 5: System Prototype User Interface.....19
Figure 6: Architecture of Simple Malay to English Translator.....21
Figure 7: Dictionary.....22
Figure 8: System Prototype Translation Accuracy in Percentage.....35
Figure 9: Sample Snapshot of Output 1.....35
Figure 10: Sample Snapshot of Output 2.....36
Figure 11: Sample Snapshot of Output 3.....36
Figure 12: Sample Snapshot of Output 4.....37
Figure 13: Project Timeline.....43

LIST OF TABLES

Table 1: List of part of Speech.....16
Table 2: Malay Sentence Structure.....17
Table 3: Test Results of Each Function.....25
Table 4: Tree Diagram of Malay Sentence Structures That Can Be Translated.....27
Table 5: System Prototype Translation Accuracy Calculation for Sentences Translated Grammatically Correct.....34
Table 6: System Prototype Translation Accuracy Calculation for Sentences Translated Grammatically Incorrect.....34
Table 7: Abbreviation for Malay Sentence Structure.....50
Table 8: Abbreviation for English Sentence Structure.....50

LIST OF FIGURES

Figure 1: Pola 1.....17
Figure 2: Pola 2.....18
Figure 3: Pola 3.....18
Figure 4: Pola 4.....18
Figure 5: System Prototype User Interface.....19
Figure 6: Architecture of Simple Malay to English Translator.....21
Figure 7: Dictionary.....22
Figure 8: System Prototype Translation Accuracy in Percentage.....35
Figure 9: Sample Snapshot of Output 1.....35
Figure 10: Sample Snapshot of Output 2.....36
Figure 11: Sample Snapshot of Output 3.....36
Figure 12: Sample Snapshot of Output 4.....37
Figure 13: Project Timeline.....43

LIST OF TABLES

Table 1: List of part of Speech.....16
Table 2: Malay Sentence Structure.....17
Table 3: Test Results of Each Function.....25
Table 4: Tree Diagram of Malay Sentence Structures That Can Be Translated.....27
Table 5: System Prototype Translation Accuracy Calculation for Sentences Translated Grammatically Correct.....34
Table 6: System Prototype Translation Accuracy Calculation for Sentences Translated Grammatically Incorrect.....34
Table 7: Abbreviation for Malay Sentence Structure.....50
Table 8: Abbreviation for English Sentence Structure.....50

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND OF STUDY

The study of natural language has been an important area of artificial intelligence almost since the beginning of the field. Two main goals motivate Artificial Intelligence (AI) work on natural language. One is the theoretical goal, and close to that of the linguist, namely, to discover how we use language to communicate. The other is technological goal, to enable the intelligent computer interfaces of the future, where natural language becomes an important means for man-machine interaction. Luckily, progress toward one of these goals often is progress toward the other a better theoretical understanding leads to more robust systems, and a better understanding of processing issues in actual applications suggests new goals and techniques of theoretical interest.

Natural language processing (NLP) can be defined, in a very general way, as the discipline having as its ultimate, very ambitious goal that of enabling people to interact with machines using their "natural" faculties and skills. This means, in practice, that machines should be able to understand spoken or written sentences constructed according to the rules of some natural language, and should be capable of generating in reply meaningful sentence in this language. The task of natural language processing (NLP) is that of accepting inputs in a human natural language, and to transform the inputs into some sort of formal statements that are to be "meaningful" for a computer.

People who need documents translated often ask themselves whether they could use a computer to do the job. When a computer translates an entire document automatically and then presents it to a human, the process is called machine translation (MT). When a human composes a translation, perhaps calling on a computer for assistance in specific

tasks such as looking up specialized words and expressions in a dictionary, the process is called human translation.

In the modern world, there is an increased need for language translation. Attempts of language translation are as old as computer themselves. Machine translation is the attempt to automate all, or part of the process of translating from one human language to another.

Machine Translation involves translating from a source natural language to a target language. Machine Translation is hard because structures in one human language often do not correspond in a simple way to structures in another. Processing natural language processing such as English has always been one of the central research issues of artificial intelligence, both because of the key role language plays in human intelligence and because of the wealth of potential applications. This project's aim is to translate simple Malay sentence to English sentence since there is not many Malay-English translator available.

1.2 PROBLEM STATEMENT

The available Malay to English translator only either translates one word a time or translates it with syntax error. The available translator prone to perform direct translation or example based translation rather then using rule based translation.

In these days of global communication and markets, there is a huge demand for translation internationally for business documents. The significant of building this translator is as the first step to building an accurate machine translator for Malay to English translation using the rule transfer translation method.

1.3 OBJECTIVES AND SCOPE OF STUDY

Language is the fundamental means of communication for human beings. The focus of this project is to build simple Malay to English Translator. There is Malay to English Translator available, but it does not translate grammatically correct. The aim of this project is to translate simple Malay sentence to English with syntactically correct.

CHAPTER 2
LITERATURE REVIEW

Natural Language Processing is the capacity of a computer to "understand" natural language text at a level that allows meaningful interaction between the computer and a person working in a particular application domain.

Natural Language Processing is the engineering of the system that process or analyze written or spoken natural language. Since most human knowledge is recorded in linguistic form, enabling computers to understand natural language will allow it to access all this knowledge. [6]

2.1 MACHINE TRANSLATION

The translation of natural languages by machine, first dreamt of in the seventeenth century, has become a reality in the late twentieth. Computer programs are producing translations - not perfect translations, for that is an ideal to which no human translator can aspire; nor translations of literary texts, for the subtleties and nuances of poetry are beyond computational analysis; but translations of technical manuals, scientific documents, commercial prospectuses, administrative memoranda, medical reports. Machine translation is not primarily an area of abstract intellectual inquiry but the application of computer and language sciences to the development of systems answering practical needs. [7]

2.2 RELATED WORK

Tong Loong-Cheong wrote a paper that presents the result obtained by English to Malay computer translation system at the level of laboratory prototype. The English to Malay translation system consists of 3 main dictionaries – source English, English-Malay transfer, target Malay- and five grammar models which are morphological analysis,

structural analysis, structural transfer, structural generation and morphological generation. [2]

2.3 TRANSLATION PROBLEM

The two main problems at the transfer phase are the incomplete (or incorrect) choice of target lexical, and the transfer of idiomatic expressions. [2]

The disambiguation of a source lexical which carries more than one meaning and which carry more than one meaning and which is translated by different target lexical accounts for more than half of the errors at transfer. The source of this problem is actually at the analysis phase, which was unable to produce a sufficiently deep level of interpretation (example: semantics and semantic relations) to solve the ambiguity which manifests itself only at transfer. [2]

Translating idioms is one of the most difficult tasks for human translators and translation machines alike. Idioms are defined as multiword expressions with a fixed (often metaphorical) meaning that cannot be derived from its parts. It is one of the most frequently used means of non-literal language. [8]

Many factors contribute to the difficulty of machine translation, including words with multiple meanings, sentences with multiple grammatical structures, uncertainty about what a pronoun refers to, and other problems of grammar. But two common misunderstandings make translation seem altogether simpler than it is. First, translation is not primarily a linguistic operation, and second, translation is not an operation that preserves meaning. [9]

2.4 SOURCE LANGUAGE ANALYSIS

A comprehensive system of natural language analysis, such as an analysis module of a knowledge-based machine translation system, must include the following basic components: [9]

- **Morphological analysis:** the decomposition of words into their uninflected root forms, performed at the word level. There are many morphological phenomena: almost all language has inflectional morphology; the majority has some form of derivational morphology. A number of general models of morphological processing have been investigated. At the theoretical level, the most popular approach to morphology is the so-called two-level approach. In practical systems many other, less general and more language and task-specific approaches have been used. [9]

- **Syntactic analysis:** the extraction of all well-formed syntactic structures and dependencies for a source text, performed at the sentence level. In the Machine Translation environment, a grammar must be written for each source language, in one of the many current grammar formalisms, such as, for instance, Lexical Functional Grammar, Generalized Phrase Structure Grammar, Head-driven Phrase Structure Grammar, Definite Clause Grammar, Tree adjoining Grammar or Government-and-Binding-related Grammars. The use of a "canonical" formalism facilitates the use of a single grammar interpreter applicable to any language whose grammar is defined in the selected formalism. [9]

- **Semantic analysis:** the creation of the knowledge structures in a text-meaning representation language (interlingua in Machine Translation) that reflect the meanings of lexical units in the source text and semantic dependencies among them, performed at the sentence level but often having to take into account suprasentential contexts. Semantic analysis procedures are typically developed for a particular domain (example: medicine, finance, and computers), though general,

"common sense" semantic knowledge is also used. The existence of canonical formalisms for encoding world knowledge and text meaning enables the use of a single universal semantic interpreter with different knowledge source for each domain. [9]

- **Pragmatic or discourse analysis:** suprasentential analysis leading to the resolution of anaphors, ellided phrases, deixis, as well as the attribution of intent and speech acts. In its full form, discourse analysis leads to the creation of a text-meaning structure in a representation language with the various domain oriented and rhetorical relations among the elements of a text, including coreference of noun phrases and anaphors, causal and temporal relations, topic/comment structure and so forth. The state of the art in pragmatic and discourse analysis is not as well developed as the other three phases of language analysis. [9]

The main word classes in Bahasa Melayu are nominal, verbal, auxiliaries, adverbials and particles and these again can be subcategorized. The word class of a derived word is dependent on its affix and on affix deletion to obtain the root form; its word class is set. For multiple affixations, the outermost prefix (if any) determines the word class. [14]

2.5 TRANSLATION METHOD

The input is separated into tense and modal information (the 'subjective' part) and the kernel sentence (the 'objective' part). The objective part is translated using the multi-level transfer method: wide ranging rules such as direct parse tree transfer are applied first, followed by idiomatic expressions then patterns from the semantic valency dictionary, and finally the default general patterns. This allows transfers of varying granulity. [3]

This process of translation can be divided into 7 parts. First, splits the input text into morphemes. Second, analyze the sentence syntactically, often giving multiple possible interpretations. Next, it rewrites the complicated source language expression into more

easily translated ones. Fourth, it semantically evaluates the various interpretations. Fifth, syntactic and semantic criteria are used to select the best interpretation. Sixth, this interpretation is used as input to generate Malay. Finally the sentence is adjusted to give the correct inflectional form. [3]

The dictionaries provide detailed information about the meanings and use of the word. Most words can express several concepts. Which concept a word expresses in a text is determined by the relationship with the other word in the text. [3]

Knowledge based system rely on the programmers to enter various languages vocabulary and syntax information into databases. The programmers then write lists of rules that describe the possible relationships among a language's part of speech. [1]

Rather then using the knowledge based system's direct word by word translation technique, statistical approaches translate documents by statistically analyzing entire phrases and, over time, 'learning' how various languages work. [1]

The translation process

Analysis is carried out in several steps; this is due to the two languages of different origins, and this renders the machine translation process more complex. The steps of the analysis: Segmentation, Morphological analysis and Linguistic units' translation. [15]

2.6 TRANSLATION APPROACHES

Example Based Approach

The example-based approach has become a common technique for Natural Language Processing applications, especially in Machine Translation. However, a main problem normally arise in the current approaches which indirectly limits their applications in the development of a large scale and practical example-based system, for example the lack of

flexibility in creating the representation tree due to the restriction that correspondences between nodes (terminal and non terminal) of the representation tree and words of the sentence must be one-to-one and some even restrict it to only in projective manner according to certain traversal order. This restriction normally results to the inefficient usage of the example base. [10]

Ideally if the sentence is already in the example-base, the translation is found there too, but in most cases, the *source* sentence will not be found in the example-base. In such case, a method is used to retrieve close related examples and use the knowledge from these examples to construct the translation for the *source* sentence. In general, this approach relies on the assumption that if two *source* sentences are "close", their translations should be "close" too; if the translation of the first one is known, the translation of the other can be obtained by making some modifications in the translation of the first one. The example-based approach has become a common technique for Natural Language Processing applications, especially in Machine Translation. However, a main problem normally arises in the current approaches which indirectly limits their applications in the development of a large scale and practical example-based Machine Translation system, i.e. the lack of flexibility in representing translation relations between *source* and *target* substrings where the substrings being possibly discontinuous in both cases. [12]

Statistical Approach

Statistical methods have proven their value in automatic speech recognition and have recently been applied to lexicography and to natural language processing. A statistical translation system requires a method for computing language model probabilities, a method for computing translation probabilities, and, finally, a method for searching among possible source sentences for the one that gives the greatest value .[11]

Direct Translation Approach

The main idea behind transformed into output (target language) sentences by carrying out the simplest possible parse, replacing source words with their target language equivalents as specified in a bilingual dictionary, and then roughly re-arranging their order to suit the rules of the target language.[9]

Transfer Based Approach

The idea of transfer-based method is clear that if the system is translating from source language to target language, the first (analysis) step involves using the parser and the source grammar to analyze the input. The second (transfer) step involves changing the underlying representation of the source sentence into an underlying representation of the target sentence. The third (synthesis) step and final major step involves changing the underlying target representation into the target sentence, using a generator and the target grammar. [9]

Corpus Based Approach

In machine translation , one major new line of investigation that has emerged in recent years is the corpus-based method which uses corpora of parallel texts in the construction of translation and computer aided translation systems to improve the translation output. [13]

Pattern Based Approach

Pattern-based machine translation systems can be easily customized by adding new patterns. To gain full profits from this character, input of patterns should be both expressive and simple to understand. The pattern-based machine translation system we have developed simplifies the handling of features in patterns by allowing sharing constraints between non-terminal symbols, and implementing an automated scheme of

feature inheritance between syntactic classes. To avoid conflicts inherent to the pattern-based approach the system has priority control between patterns and between dictionaries. [18]

2.7 DESIGN OF THE BILINGUAL DICTIONARY

There will be 7 fields in the dictionary which are Malay Index Word , Malay Root Word , Part-of-Speech Noun , Syntactic Features , Semantic Features , English Translation , and English Definition .[4]

Malay Root Word

If the index word is a derived form, then the root is entered here. [4]

Malay Part-of-Speech

This is the part-of-speech of the Malay index word. We have a hierarchy of parts-of-speech. There is an initial split into 5 groups: nouns, verbs, adjectives, adverbs and function words. Within each word class, further details are given, for example the noun word class is further divided into five distinct types: common nouns, proper nouns, numeral classifiers, pronouns and titles. [4]

Parts of speech are labeled with a three-letter code. All the noun codes begin with N to indicate that the word/phrase in field one is from the class of noun. For ease of processing, codes (the second and third letters) from each of the five distinct types begin with different letters. The following two letters in each code will indicate what type of noun index word is, e.g., NC refers to common noun or NP refers to proper noun. Hence the part-of-speech codes are NNC for common nouns and NNP for proper nouns. Altogether we distinguish 38 different parts of speech. [4]

Syntactic Features

More detailed syntactic information is given in this field, either as single codes, or codes with values.Some examples of codes are given here. NHR is used to show a multi-word noun with the head on the right (the default in Malay is for the head to be on the left). CL is used to mark the default classifier associated with each noun. This code must be accompanied by a value, which is the classifier, given as a string. Malay is a numeral classifier language, and nouns normally cannot be directly modified by numerals. Instead a classifier is used, similar to *piece* in *two pieces of paper*. The classifier is useful for generation in NLP applications, and also for second language learners. Classifiers are listed for 18,900 Malay nouns (22%). We are trying to put in the minimum amount of syntactic information needed to correctly generate Malay. [4]

Semantic Classes

Semantic information is stored in this field. The most common information is semantic classes from the Goi-Taikei ontology (Ikehara *et al.*, 1997). This is a hierarchical ontology of some 3,000 semantic classes, organized with is-a and has-a relations. The classes can be used to disambiguate words with multiple senses. For example *perang* has two distinctive meanings: "brown" (as in colour) and "war". .[4]

Separate entries will be created for these senses, the first marked with [2352:color] and the second with [1755:war]. In some cases we wished to distinguish semantic classes not given in the Goi-Taikei ontology. For example, [bladed weapon] is a natural class in Malay, and has its own classifier *bilah*. In such cases we have added semantic classes to the ontology (so far we have only added [bladed-weapon] and [substance]). [4]

English Translation

This field lists one or more translation equivalents for the Malay index word or phrase. If there is more than one translation, and using different semantic classes cannot

disambiguate the translations, then they will all be listed, with the translation judged to be most common given first. For example , *gerombolan* can be translated as 'band', 'gang' or 'group'. Because *group* is the most general translation, we list it first. Of course, multiple Malay words may have the same English translation: e.g. *pencernaan* and *penghadaman*, both have 'digestion' as their English translation. .[4]

English Definition

When we can't immediately determine a good translation equivalent, we add a definition in this field. For example, many plant and animal names have no exact equivalent. For these, we leave the translation field blank, and translate as the Malay word in single quotes, followed by the definition in brackets: for example, *mempinas* (a kind of fish), *mempitas* (a kind of tree). [4]

2.8 ISSUES IN THE CONSTRUCTION OF THE BILINGUAL DICTIONARY

According to Chiew Kin Quah, Francis Bond and Takefumi Yamazako,
the base of our Malay-English lexicon was a dictionary produced by a Malaysian translation company. This included English, Malay and Chinese, with some part-of-speech information and numeral Classifiers. We reformatted the dictionary so that each entry had the nine fields described above, plus an ID field and a field for comments by the lexicographer. The entire dictionary is kept as a single text file. It is around 8 Mega bytes, which is small enough to edit using the text editor Emacs, which we use in both Windows and UNIX environments. During construction, we used a variety of one-off Perl scripts to add, change and reformat information automatically. [4]

It is possible to create a useful bilingual dictionary by matching two or existing bilingual dictionaries and linking them through a pivot language. The accuracy of the resulting dictionary can be improved by (1) using semantic classes, and (2) matching through a different pivot language. [5]

Semantic Dictionary

In order to realize semantic analysis, use semantic dictionaries and structure dictionaries, dictionaries based on a precise and detailed semantic attribute system. The lexicons provide detailed information about the use and meanings of words. Most words can express several concepts. Which concept a word expresses in a text is determined by its relationship with the other words in the sentence. For example the noun "hotel" is used as "a common noun - concrete - an agent - an organization - a corporation" and as "a common noun - concrete - a place - a facility - a lodging facility". [16]

There are a lot of different bilingual dictionaries available in the world. Still, that does not mean that there is a bilingual dictionary for every pair of languages. Considering two 'minor' languages like Malay Indonesian and Hungarian, there is a very slim chance that a dictionary translating between these two languages will be found. This is not surprising since there are several thousand different languages in the world today, so a full coverage would require many millions of bilingual dictionaries. [17]

CHAPTER 3

METHODOLOGY/PROJECT WORK

The methodology that is applied through out the system is System Development Life Cycle (SDLC). The methodology consists of Planning, Analysis, Design, Development and Testing as the guide and a framework for completing the project.

3.1 METHODOLOGY

3.1.1 PLANNING

In the planning phase, the projects' problem statement, objectives, scope and schedule are established. The purpose of this phase is to define and establish the project plan emphasizing the important thing such as the purpose of the project, the significant of developing the project, and the due date of the project. The activities conducted during the planning phase can be referred to Appendix I in the Appendices section.

3.1.2 ANALYSIS

Analysis phase consists of activities such as analyze the current problem and technology. Vast amount of journals and articles are used as reference on previous work conducted by other researchers. Research is done regarding the work that have been done on machine translation, the approach used, the current situation and problem of machine translation. Research is done on how to build the dictionary, what transfer method to be use in the translation. During building the dictionary itself , it consume quite some time since when people think of storing dictionary in proper database such as Microsoft Access etc. After some reading, it is proven that for a machine translation, the dictionary should be kept in a single text file. There are many translation methods available such as Direct Translation, Semantic Translation, Rule Based Translation, Example Based Translation and etc. For this project, the Rule Based Translation approach will be used.

The list of part of speech in Malay sentence structure:

Table 1: List of part of speech

	Malay	English
1.	Angka	Numeral
2.	Kata adjektif	Adjective
3.	Kata adverba	Adverb
4.	Kata arah	Directional noun
5.	Kata bantu	Auxiliary
6.	Kata bilangan	Numeral
7.	Kata ganti nama	Pronoun
8.	Kata hubung	Conjunction
9.	Kata kerja	Verb
10.	Kata naïf	Negative
11.	Kata nama	Noun
12.	Kata penegas	Emphatic word
13.	Kata penguat	Intensifier
14.	Kata perintah	Imperative word
15.	Kata sandang tentu	Definite article
16.	Kata sandang tak tentu	Indefinite article
17.	Kata sendi	Preposition
18.	Kata seruan	Interjection
19.	Kata Tanya	Interrogative word
20.	Penjodoh bilangan	Numeral classifier
21.	Singkatan	Abbreviation
22.	Jamak	Plural
23.	Tunggal	Singular
24.	Kata penunjuk	Determiner
25.	Kata ganti nama diri	Personal pronoun

26.	Kata ganti nama tunjuk	Demonstrative pronoun
27.	Kata ganti nama tak tentu	Indefinite pronouns
28.	Adverbial	Adverbials
29.	Kata kerja bantu	Auxiliary verbs
30.	Kata kerja modal	Modal verbs
31.	Kata kerja refleksif	Reflexive verbs

For Bahasa Melayu, it has 4 main sentence structures compared to English which only has one main sentence structure which is ***Noun Phrase + Verb Phrase***. The Malay main sentence structures are:

Table 2: Malay sentence structure

Pola Ayat	Subjek	Predikat
Pola 1	Frasa Nama (Ali	Frasa Nama guru)
Pola 2	Frasa Nama (Adikku	Frasa Kerja sedang tidur)
Pola 3	Frasa Nama (Pegawai itu	Frasa Adjektif amat rajin)
Pola 4	Frasa Nama (Rumahnya	Frasa Sendi Nama di Kampung Baru)

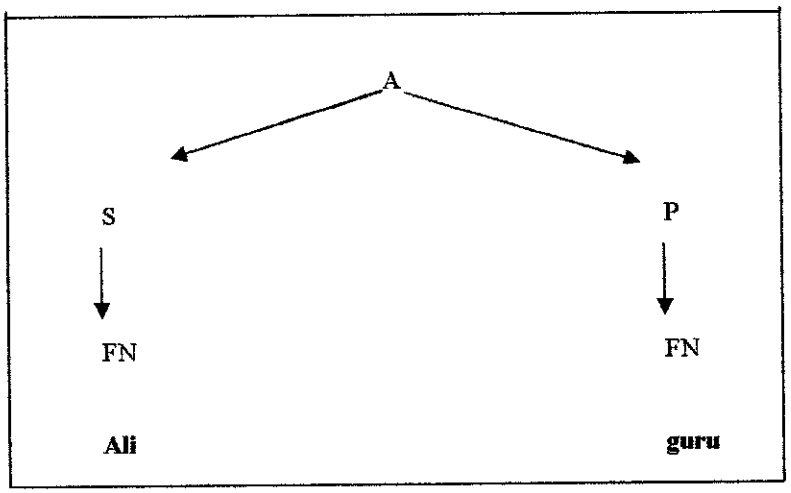


Figure1: Pola 1

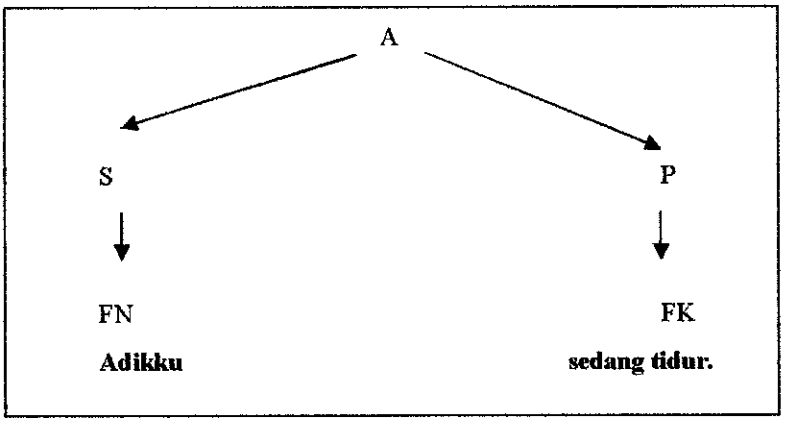


Figure2: Pola 2

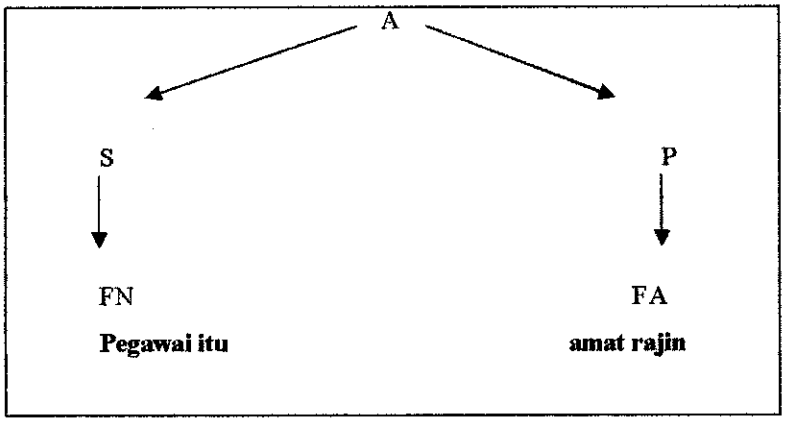


Figure3: Pola 3

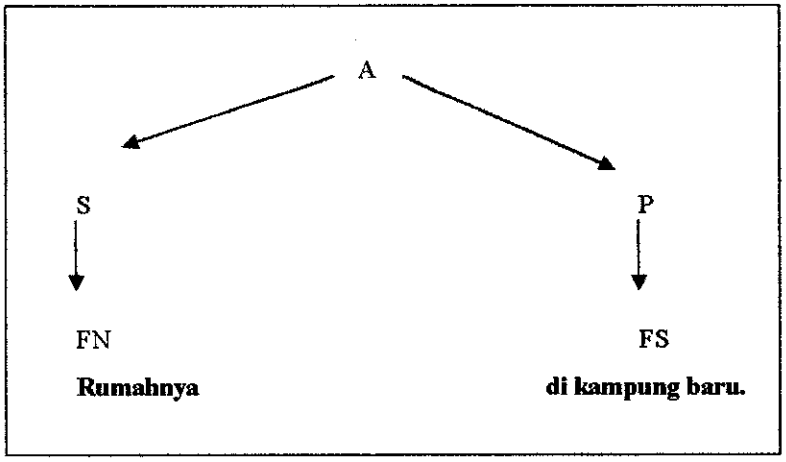


Figure 4: Pola 4

These Malay sentence structures are obtained from *Tatabahasa Dewan* by Nik Safiah Karim,Farid M. Onn, Hashim Hj. Musa, Abdul Hamid Mahmood. [19]

3.1.3 DESIGN

- Design the Simple Malay to English Translator user interface.

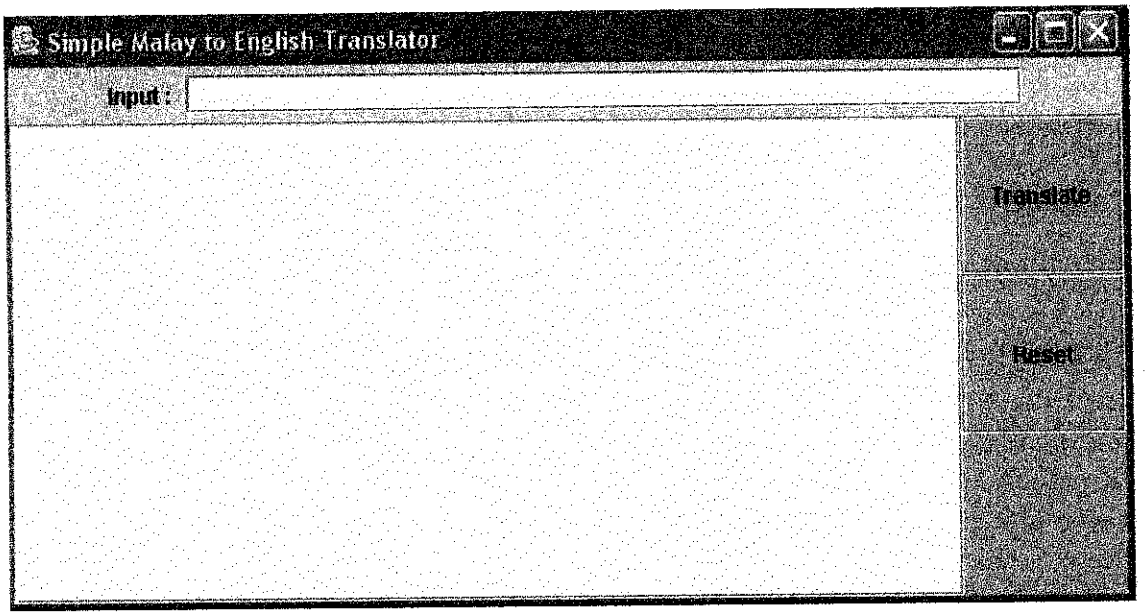


Figure 5: System Prototype User Interface

The user interface of the system prototype consists of one input field, one output text area, and two buttons. The buttons are named Translate and Reset. The interface is designed to allow user to use the system prototype easily.

3.1.4 DEVELOPMENT

This is the core programming part whereby the translator is developed using Java language. Develop the translator using Java. Try to translate simple Malay sentence to English.

Developed a package and name it Translator. Created a class and name it Dictionary. In this application, there are 5 methods which are Dictionary, FindDef, FindPOS, Translate WithRules and Clear.

3.1.5 TESTING/DELIVERED

And lastly the testing phase where any errors or bugs will be detected and corrected. Testing is done at every deliverable phase. Testing is done at sub content of the application that leads to the testing of the whole application when the application is ready.

Testing is done at every function which consists of Dictionary, FindDef, FindPOS, Clear and TranslateWithRules. This is called unit testing. Then, the prototype will be tested as a whole application. This phase is called integration testing. Finally, the project will be presented and delivered.

3.2 SIMPLE MALAY TO ENGLISH TRANSLATOR ARCHITECTURE

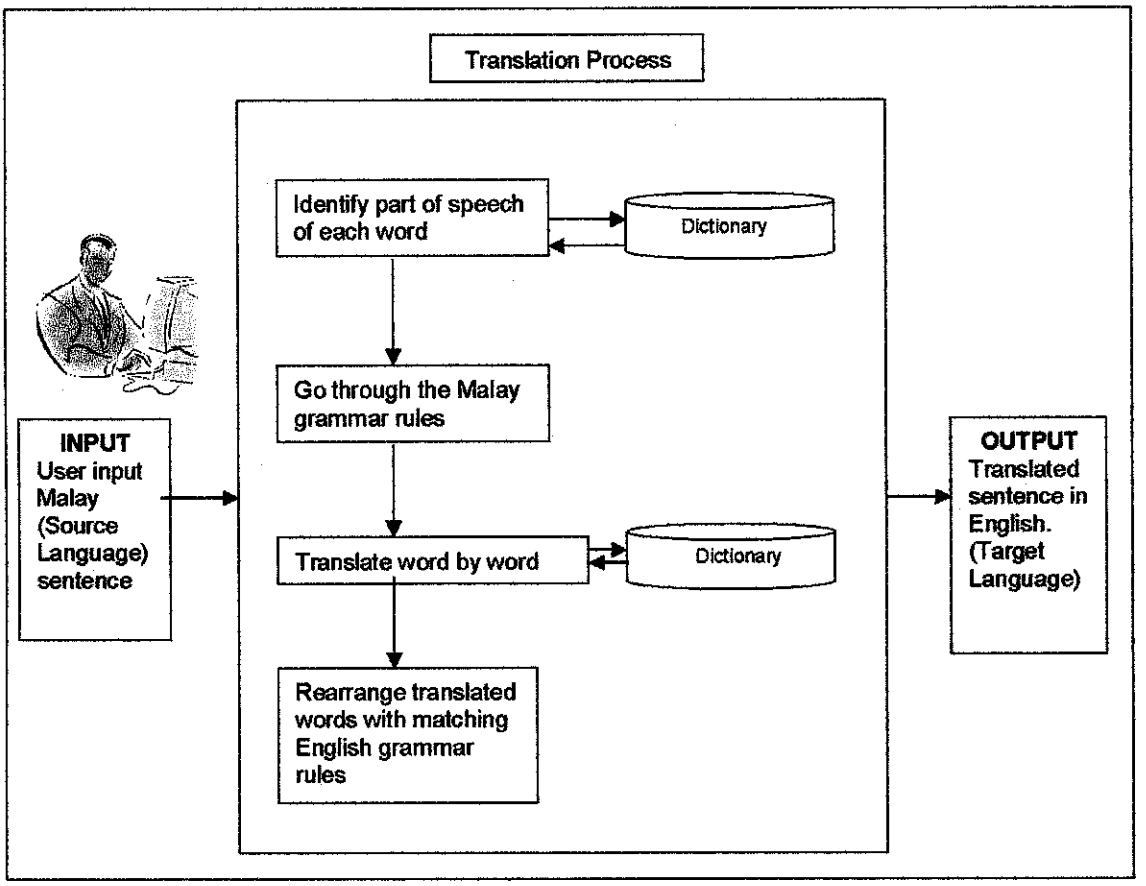


Figure 6: Architecture of Simple Malay to English Translator

Below are the explanations for the Architecture of Simple Malay to English Translator:

INPUT

User will enter the source natural language sentence which is simple Malay sentence.

TRANSLATION PROCESS

After user enters the sentence, the system prototype will identify the part of speech of each word by referring to the dictionary. In the dictionary there are 3 fields which are Malay Word, Part-Of-Speech and English Word.

Overall, there are 28 rules in the system prototype. If one of the rules is met, it will translate the sentence word by word by referring to dictionary. After translating word by word, the words will be rearranged according to the matching English grammar rules.

OUTPUT

Finally, the output is presented to user which is the translated sentence in English.

3.3 MALAY - ENGLISH DICTIONARY DATABASE

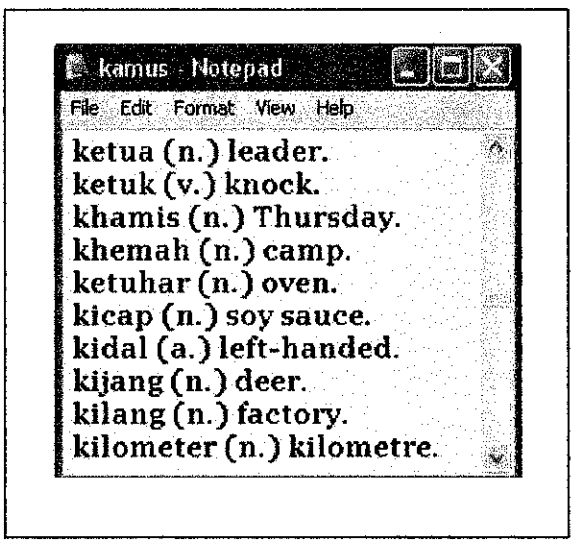


Figure 7: Dictionary

In designing the Malay-English dictionary database, the entire dictionary is kept in a single text file. From earlier research, it is proven that the dictionary for a machine translator should be kept in a single text file for easier and faster retrieval due to the small file size. In the dictionary, there are 3 fields which are Malay Word, Part-Of-Speech and English Word. For this system prototype, the Collins Headstart Bilingual Dictionary [23] is used.

This dictionary is chosen as it fits the following criteria:

- Quantity of information
- Quality of information
- Effectiveness of presentation

Overall, the dictionary contains 2120 words which were entered manually into the single text file.

3.4 TOOLS

The main tools that will be used for the project development are Java Language, Forte for Java 4.0 Community Edition and Microsoft Notepad version 5.1. The dictionary is build using the Microsoft Notepad version 5.1 and the application is developed using Java Language and Forte for Java 4.0 CE.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 RESULTS AND DISCUSSION

From earlier research, a dictionary that is used for a machine translator is usually being created in a notepad file for easy retrieval compared to using Microsoft Access or other database application.

According to Chiew Kin Quah, Francis Bond and Takefumi Yamazako,

> the base of our Malay-English lexicon was a dictionary produced by a Malaysian translation company. This included English, Malay and Chinese, with some part-of-speech information and numeral Classifiers. We reformatted the dictionary so that each entry had the nine fields described above, plus an ID field and a field for comments by the lexicographer. The entire dictionary is kept as a single text file. It is around 8 Mbytes, which is small enough to edit using the text editor Emacs, which we use in both Windows and UNIX environments. During construction, we used a variety of one-off Perl scripts to add, change and reformat information automatically.

In developing a machine translation, a logic programming language is needed in order to create the parse tree for Malay and English language. A machine translation can be developed using Java language since it has the capabilities to create logic language.

4.2 UNIT TESTING

As mentioned earlier, the system will need to through unit testing to detect any bugs in the system.

Table 3: Test Results of Each Function

Function	Expected Test Result	Actual Test Result
Clear	Able to clear the input and output field.	Successfully able to clear the fields.
FindDef	Able to retrieve definition of each word inserted.	Successfully retrieve the definition of each word inserted.
FindPOS	Able to retrieve part of speech of each word inserted.	Successfully extract the part of speech of each word from the dictionary.
Dictionary	Able to translate the inserted words.	Successfully translate the inserted words. When user inserts Malay word with spaces, it still can translate properly. Error "No Definition" appear when user insert Malay word with characters other than the full stop character '.'
TranslateWithRules	Able to translate the simple sentence inserted.	Successfully translate the simple sentence inserted based on the rules set. If the rules is not set yet, system prototype not able to translate. If the sentence structure entered does not conform the rules in the function, it will return a string at the output field stating, 'Please enter a proper sentence.'

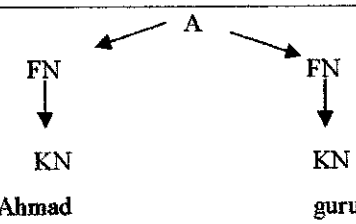
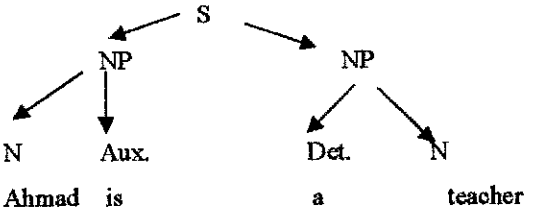
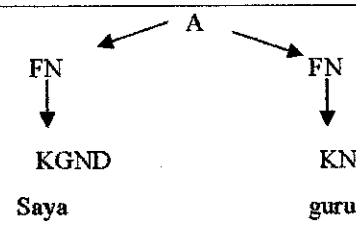
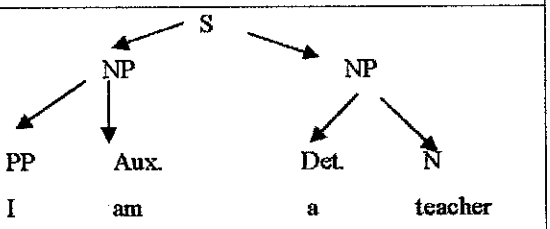
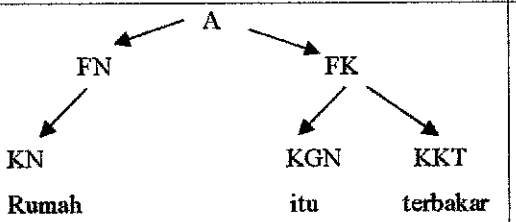
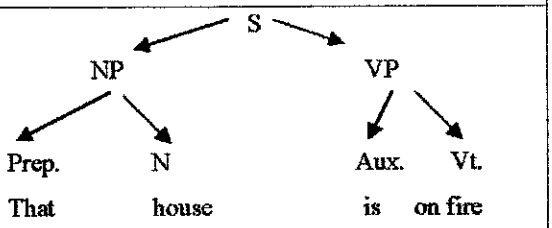
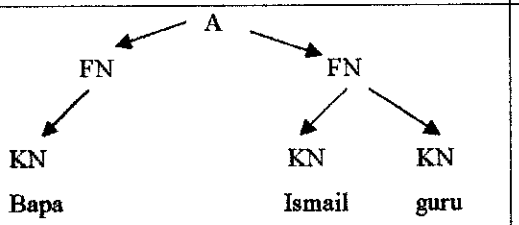
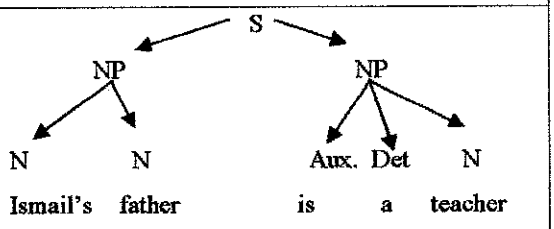
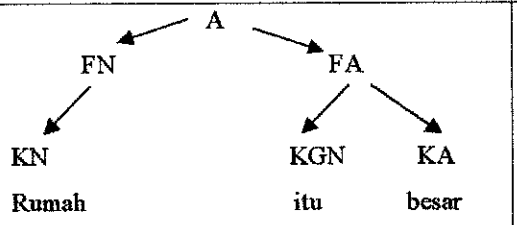
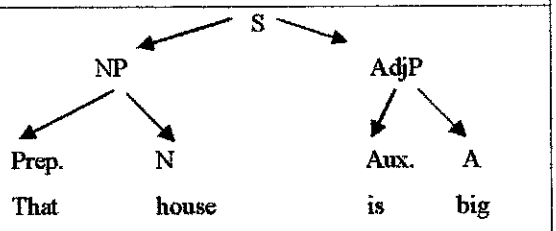
4.3 INTEGRATION TESTING

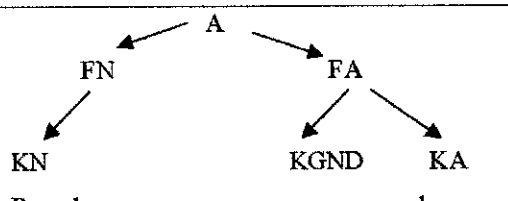
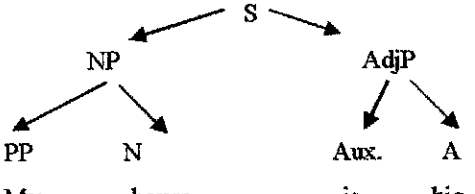
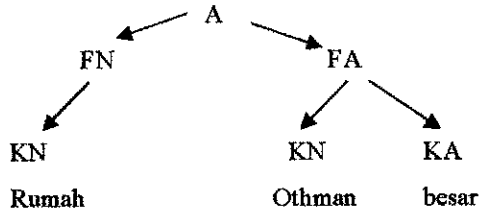
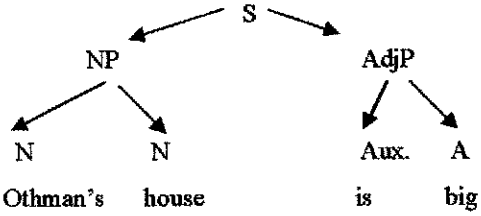
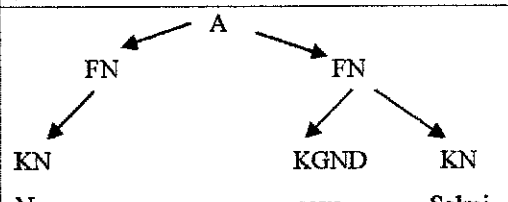
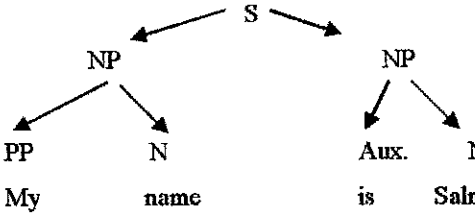
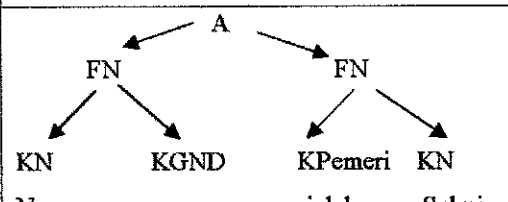
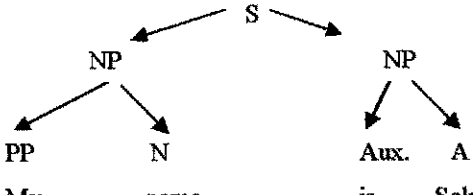
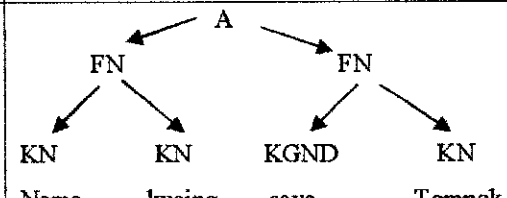
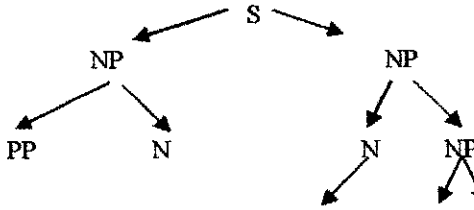
After performing the unit testing, the system prototype is tested as a whole unit. Testing the system prototype as a whole unit is called Integration Testing. After clicking the button 'Translate', the system prototype is able to translate the sentence inserted by user into English sentence. This is the main function of the translator since the translator objective is to translate simple Malay to English sentences.

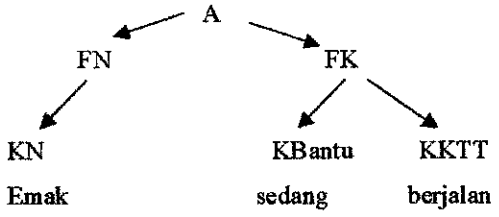
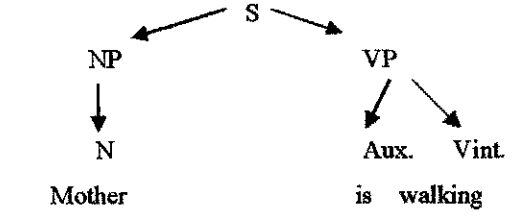
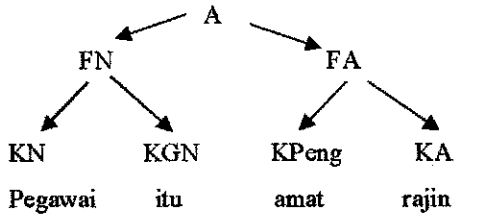
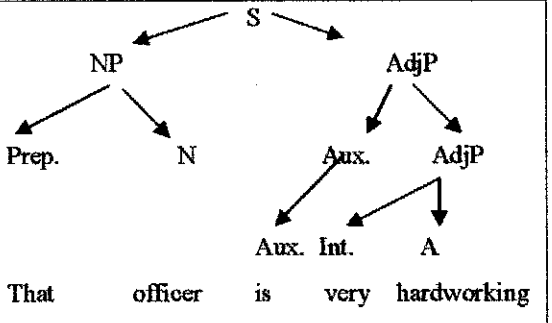
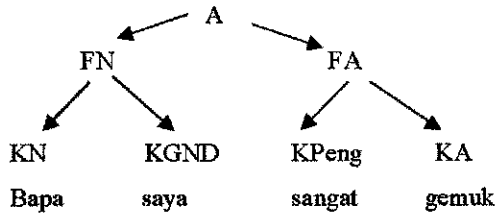
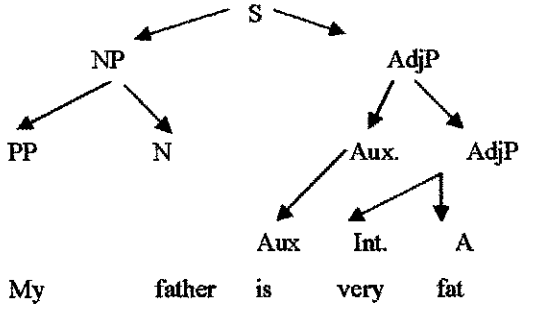
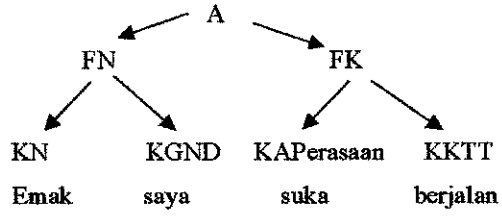
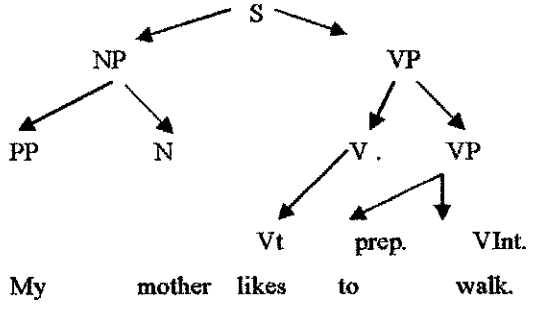
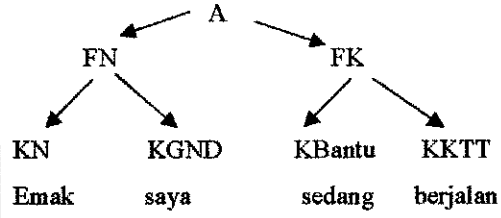
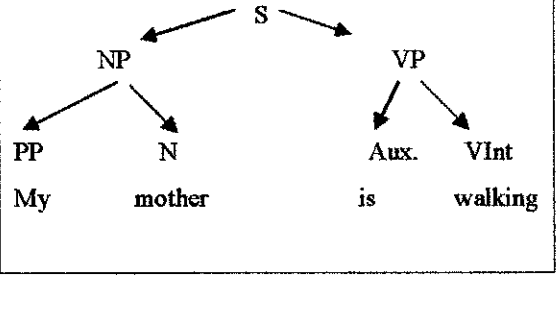
4.4 TREE DIAGRAM OF MALAY SENTENCE STRUCTURES THAT CAN BE TRANSLATED

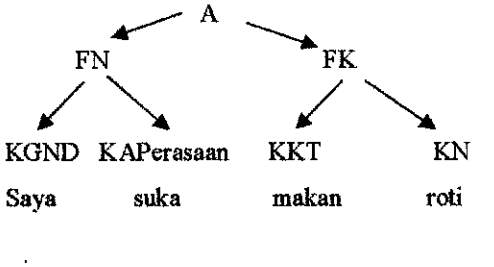
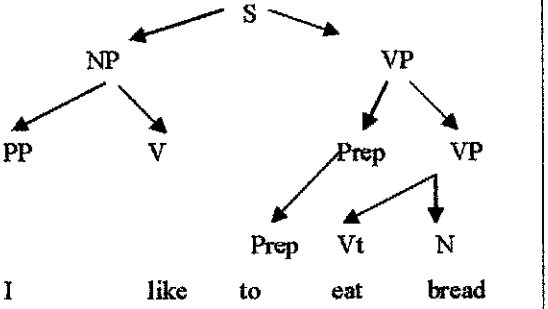
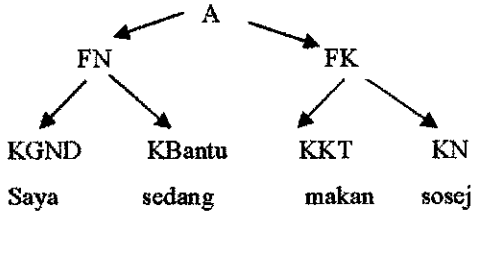
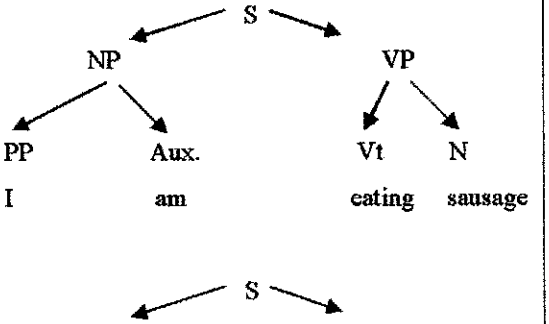
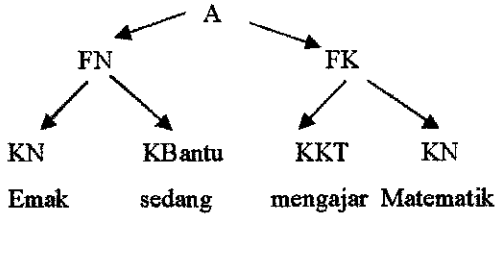
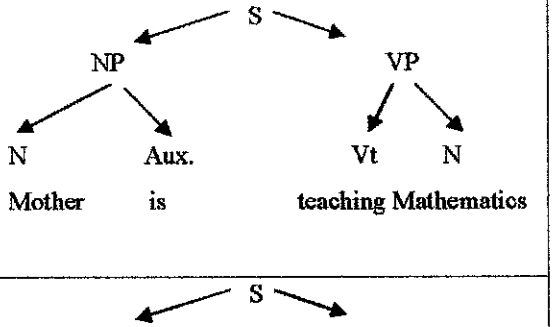
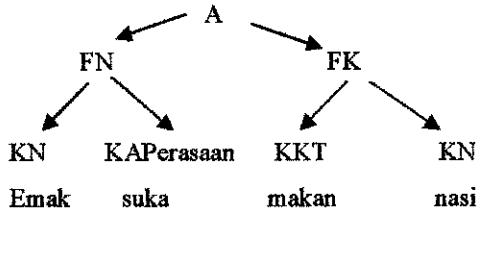
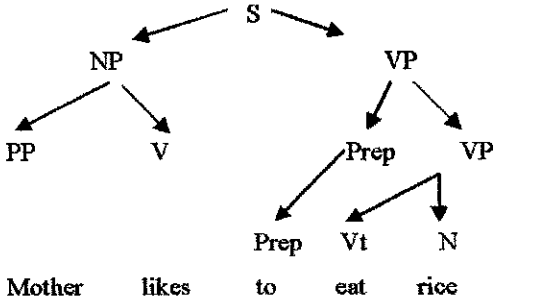
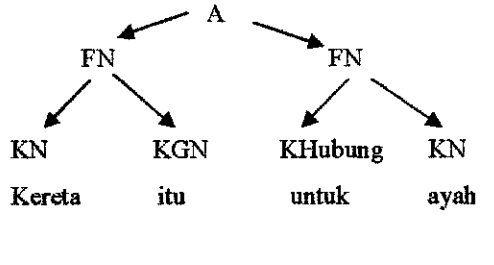
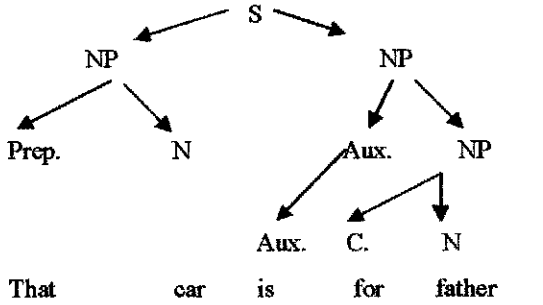
Until now, there are 28 sentence structures that can be translated. Below are the Malay sentence structure and the translated sentence structure in English:

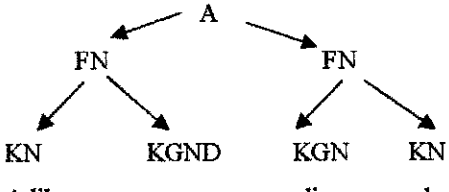
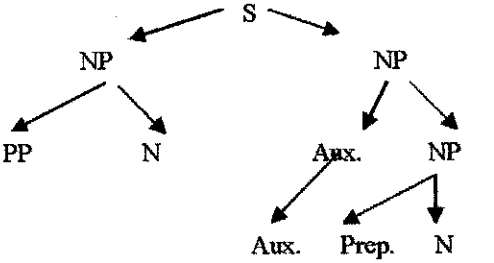
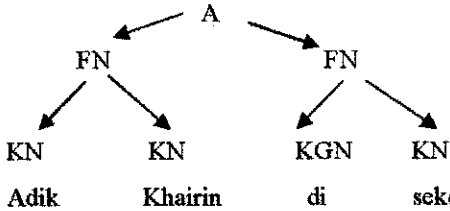
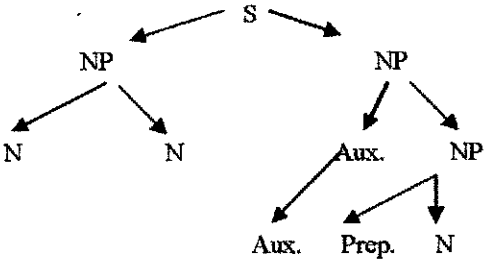
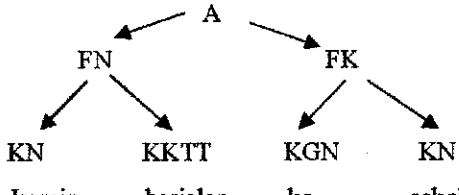
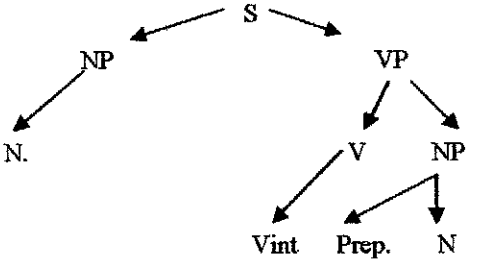
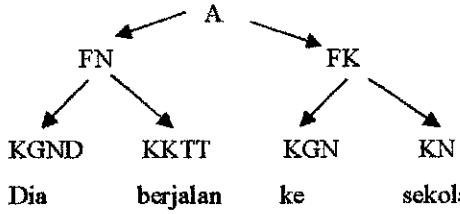
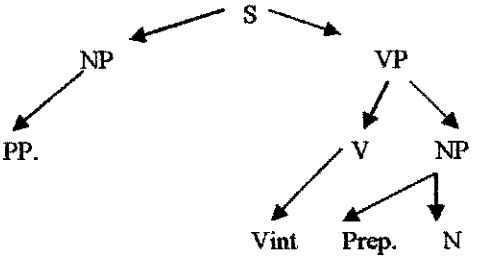
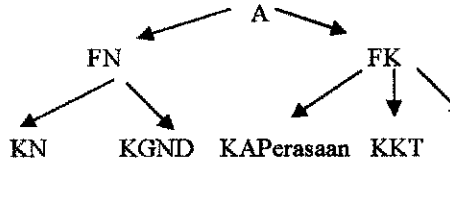
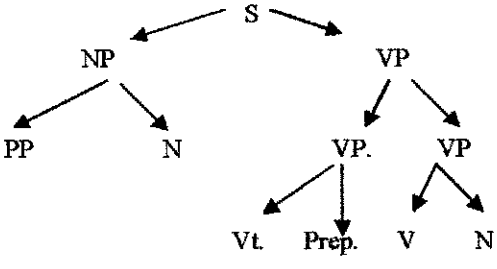
Table 4 : Tree Diagram of Malay Sentence Structures That Can Be Translated

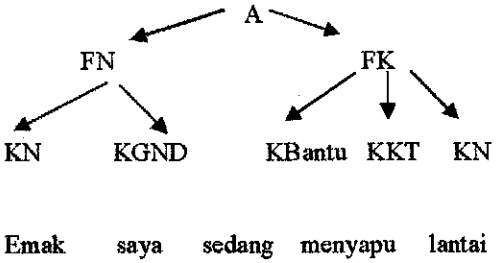
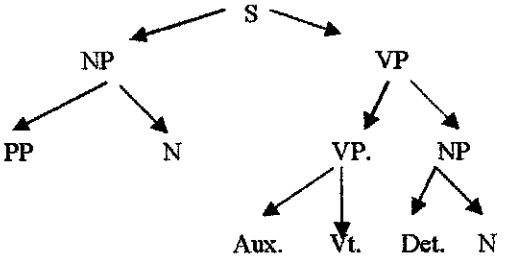
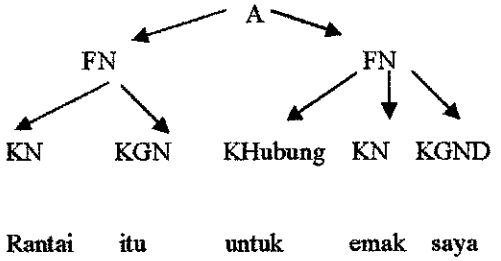
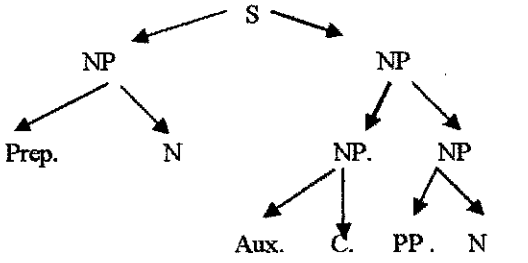
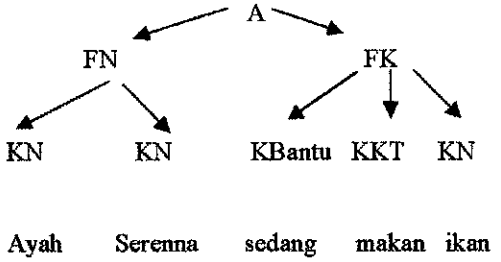
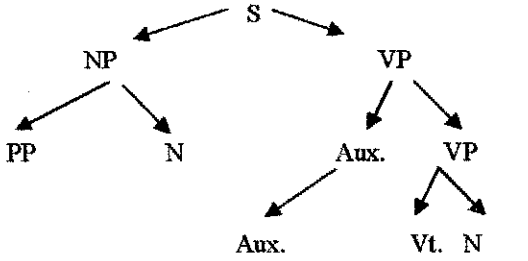
No.	Malay Structure	English Structure
1.		
2.		
3.		
4.		
5.		

No.	Malay Structure	English Structure
6.		
7.		
8.		
9.		
10.		

No.	Malay Structure	English Structure
11.		
12.		
13		
14.		
15.		

No.	Malay Structure	English Structure
16.		
17.		
18.		
19		
20.		

No.	Malay Structure	English Structure
21.		
22.		
23.		
24.		
25.		

No.	Malay Structure	English Structure
26.		
27.		
28.		

Overall , the system prototype able to translate 28 Malay sentence structures. The system prototype translates from Malay sentence to English present tense and present continuous tense only. The system prototype has not cater the translation into other English tenses such past tense and many more. The system protoype also translate Malay sentence structure that consist of Frasa Nama , Frasa Kerja and Frasa Adjektif. Refer to Appendix III for the explanation of the abbreviation used in table 4.

4.5 USER'S FEEDBACK

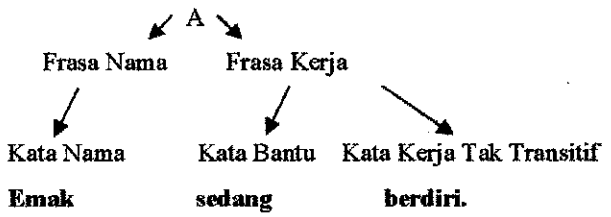
Generally, user understands the objective of the system prototype which is to translate sentence with grammatically correct. But there are users that questioned the ability of the prototype. For example , a user comment that instead of translating 'Saya sedang makan ikan' into 'I am eating fish' , the user expected that sentence to be translated to 'I am having fish'.

Some of the users commented that the translator only translate present tense and present continuous tense. They expected the system prototype able to translate most of the Malay sentence structures including the complex sentence structures.

4.6 SYSTEM PROTOTYPE TRANSLATION ACCURACY

A set of questionnaire were distributed to 10 users which consist of UTP lecturers that teach Technical & Professional Writing, Professional Communication Skills and tuition center tutors who teach English.

The sample questions from the questionnaire are as follow:



No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak sedang berdiri	Mother is standing.		
2.	Emak sedang berjalan.	Mother is walking.		
3.	Emak sedang menangis.	Mother is crying.		

Each questionnaire consists of 51 sentences and its translated sentences for all the Malay sentence structures that can be translated by the system prototype. User will need to answer whether the translated sentence is translated grammatically correct or otherwise. The questionnaires used are placed in the appendices section. There are only 2 definite answers for each question in the questionnaire which is yes or no. The questionnaire distribution is performed to evaluate the percentage of accuracy of the system prototype.

From the questionnaire distributed, it is found that 37 sentences were translated grammatically correct and 14 sentences were translated grammatically incorrect.

Table 5: System Prototype Translation Accuracy Calculation For Sentences Translated Grammatically Correct

No. of Sentences Translated With Correct Grammar (C)	Total No. of Sentences Translated (T)	Percentage of Sentences Translated With Incorrect Grammar (C/T) X 100
37	51	72.55%

From table 5, the system prototype is able to translate 72.55% grammatically correct.

Table 6: System Prototype Translation Accuracy Calculation For Sentences Translated Grammatically Incorrect

No. of Sentences Translated With Incorrect Grammar (I)	Total No. of Sentences Translated (T)	Percentage of Sentences Translated With Incorrect Grammar (I/T) X 100
14	51	27.45%

From table 6, the system prototype is able to translate 27.45 % grammatically incorrect.

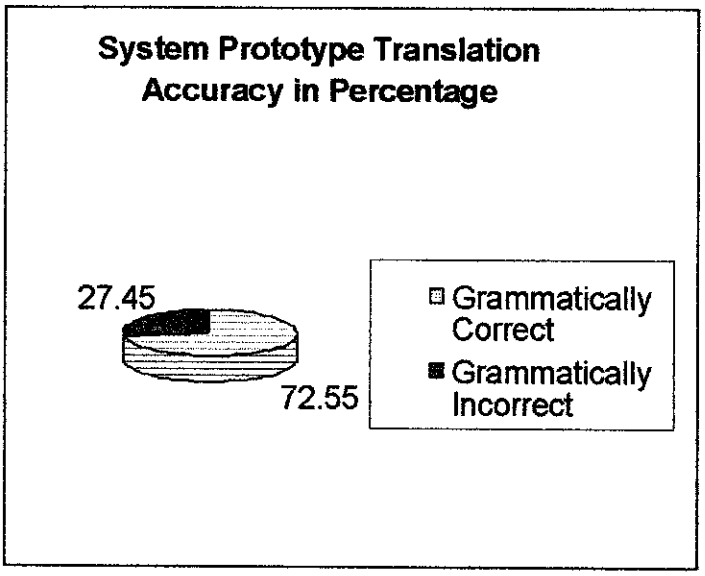


Figure 8: System Prototype Translation Accuracy in Percentage.

From figure 8, the system prototype is able to translate 72.55% grammatically correct and 27.45 % grammatically incorrect.

4.7 SAMPLE SNAPSHOTS OF OUTPUT

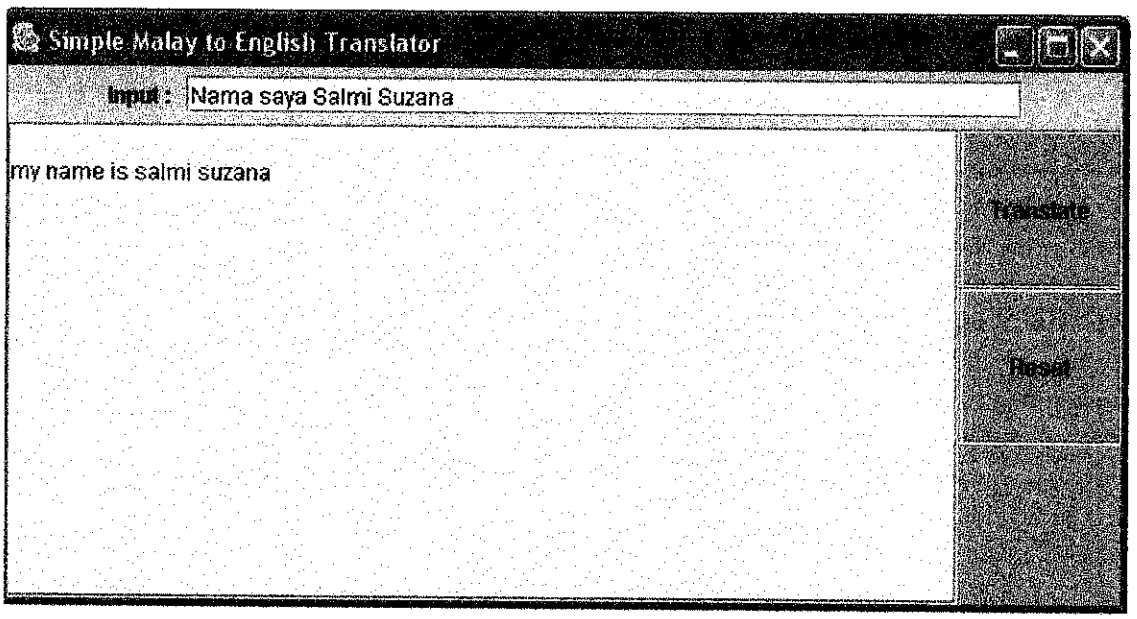


Figure 9: Sample snapshot of output 1

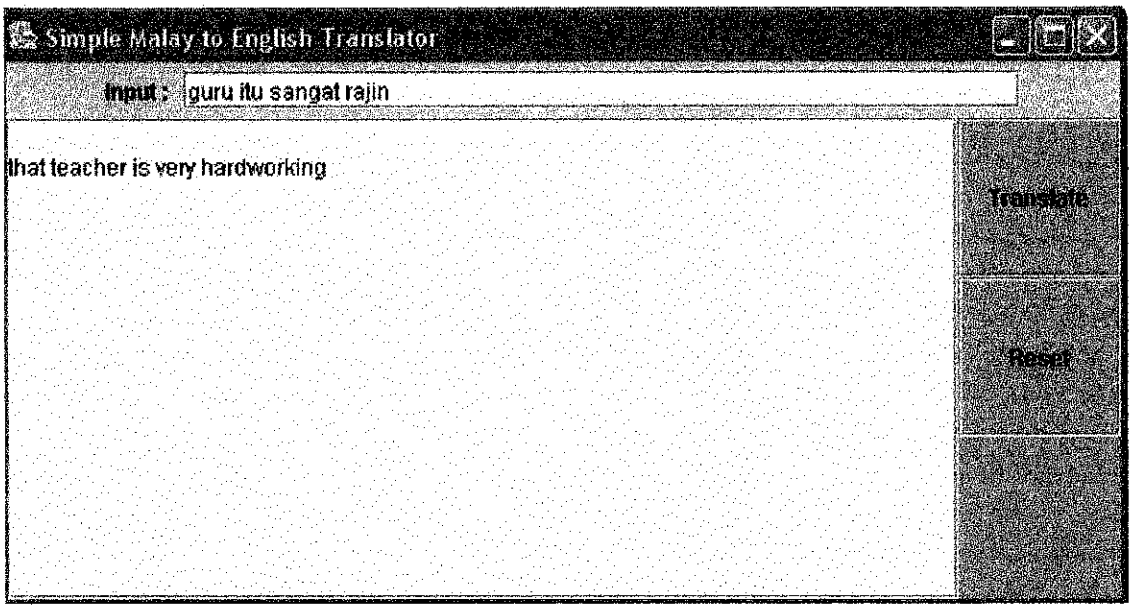


Figure 10: Sample snapshot of output 2

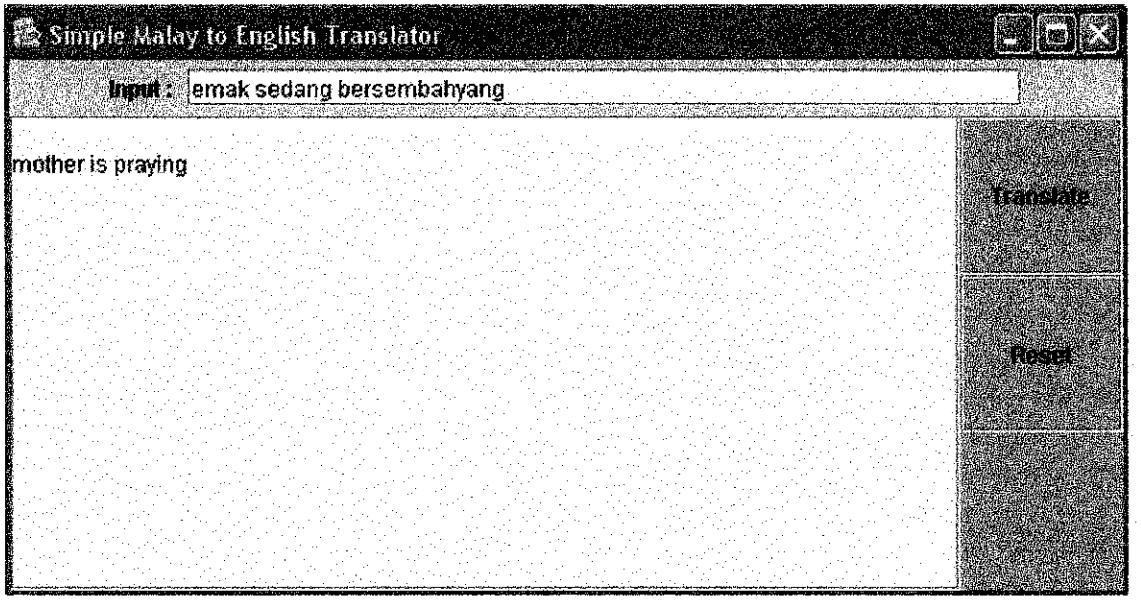


Figure 11: Sample snapshot of output 3

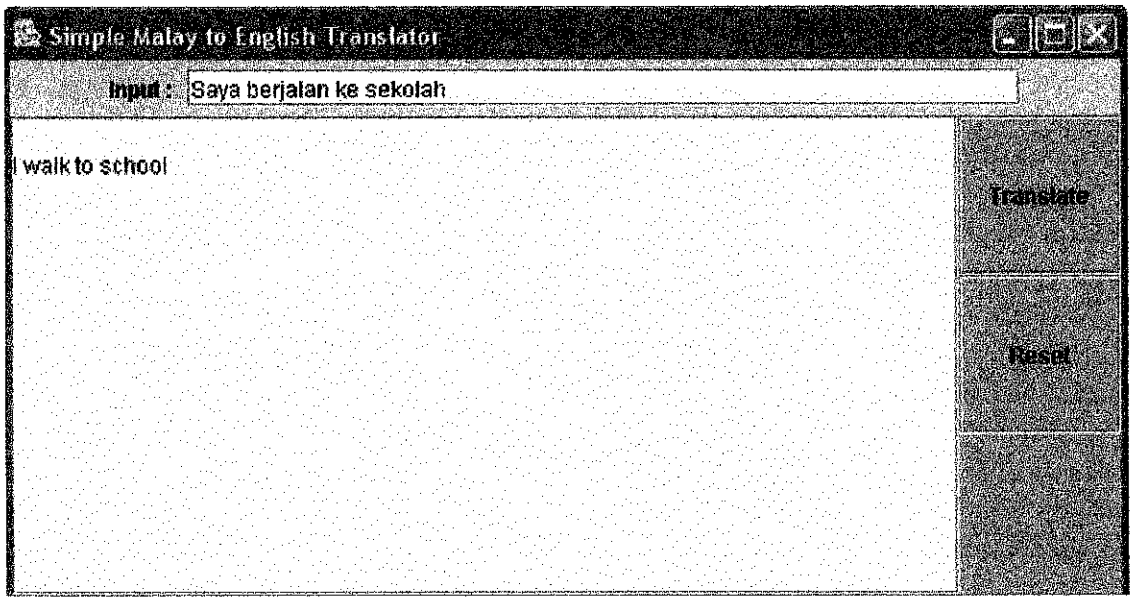


Figure 12: Sample snapshot of output 4

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

Due to the time limitation, the scope of the project is reduced to translating only simple Malay sentence to English sentence. Until now, the system prototype is able to translate 28 Malay sentence structures. The system prototype is able to translate into English tenses such as present tenses and present continuous tense only.

The ambiguity problem and the semantic factor would not be addressed in this project. The Simple Malay to English Translator has met its objective which is to translate simple Malay to English sentence.

5.2 RECOMMENDATION

- **5.2.1 DICTIONARY**
  - The dictionary should contain more words since there are many possibilities of sentence structure with different words.

- **5.2.2 SIMPLE MALAY TO ENGLISH TRANSLATOR**
  - The simple Malay to English translator should be able to translate more Malay sentence structure in future.
  - It should not only consider the syntactic issue, the semantic issue should also be taken into consideration. The translator should be able to translate with correct grammar and the translated sentence makes sense.

REFERENCES

[1] Neville Holmes, (2004, March). *Languages and Computing Profession.*

[2]]Tong Loong Cheong. *English-Malay Translation System: a Laboratory Prototype*, Proceedings of the 11th International Conference on Computational Linguistics, COLING-86, Bonn, August 1986.

[3]Kentaro Ogura, Francis Bond, Yoshifumi Ooyama. *Alt-J/M: A prototype Japanese to Malay Translation System.*

[4]Chiew Kin Quah, Francis Bond and Takefumi Yamazaki, *Design and Construction of a Machine-Tractable Malay-English Lexicon.*

[5]Francis Bond, Ruhaida Binti Sulong , Takefumi Yamazaki and Kentaro Ogura, *Design and Construction of a machine-tractable Japanese-Malay dictionary.*

[6]Tan Chin Tuck, *Natural Language Processing. Understanding what you say.*

[7] W.John Hutchins, *Machine Translation: A Brief History*

[8] Martin Volk ,*The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems*

[9] Azza Abd El-Moniem Mohamed , *Machine Translation of Noun Phrases : From English to Arabic*

[10] Tang Enya Kong , Mosleh Hmoud Adhaileh : *Example-based natural language parsing on the SSTC annotation schema.*

[11]Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek,John D. Lafferty, Robert L. Mercer, and Paul S. Roossin : *A Statistical Approach to Machine Translation.*

[12] Mosleh H. Al-Adhaileh , Tang Enya Kong : *Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema*

[13] Mosleh H. Al-Adhaileh , Tang Enya Kong : *S-SSTC: A Synchronous Annotation Schema of Structured String-Tree Correspondence (SSTC)*

[14] Chang May See , *The Morphological Analysis of Bahasa Malaysia.*

[15] *Haytham* Alsharaf , Sylviane Cardey and Peter Greenfield , *French to Arabic Machine Translation: the Specificity of Language Couples.*

[16] Minoru Hayashi, Setsuo Yamada, Akira Kataoka and Akio Yokoo : *ALT-J/C A Prototype Japanese-to-Chinese Automatic Language Translation System.*

[17] *Multilingual Lexical Database*

[18] Mihoko Kitamura and Toshiki Murata: *Practical Machine Translation System allowing Complex Patterns*

[19] Nik Safiah Karim,Farid M. Onn, Hashim Hj. Musa, Abdul Hamid Mahmood 1997 , *Tatabahasa Dewan* , Kuala Lumpur , Dewan Bahasa dan Pustaka.

[20] Wardhaugh 2002, Understanding *English Grammar A Linguistic Approach*, United Kingdom, Blackwell Publishing.

[21] Y.Danial Liang 2003, *Rapid Java Application Development Using Sun One Studio 4*, United States of America , Prentice Hall.

[22] Harvey M.Deitel and Paul J.Deitel 2005,*JAVA How To Program 6th Edition* ,United States of America ,Prentice Hall.

[23] *Collins Headstart Bilingual Dictionary* , United Kingdom, HarperCollins Publisher.

APPENDICES

Appendix I: Project Timeline

Figure 13: Project Timeline

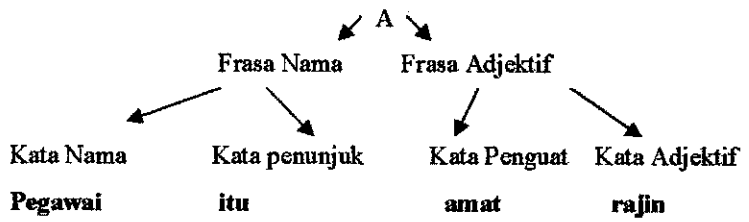
ACTIVITIES	WEEK NO/DATE															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Research of Proposal (Malay to English Translator)																
FYP Briefing Understand NLP concept Understand Malay and English Parsing																
Research Work - Distribution of questionnaire (if any)																
System Analysis																
System Design																
System Development																
System Testing																
Revision of System																
Preparation on Final Report/ Dissertation																

Appendix II: Questionnaire

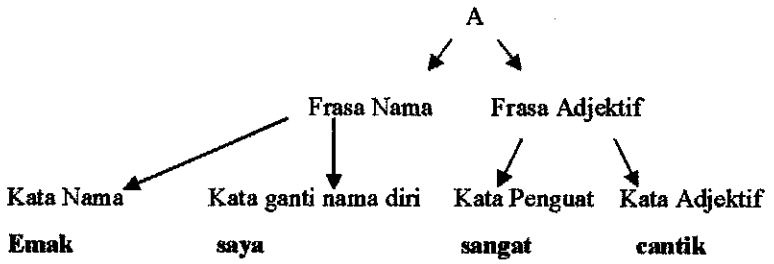
Evaluation for Simple Malay to English Translator

The following are the questionnaires related to Simple Malay to English Translator. Please provide your answer by tick in the box that applies. Please answer all questions.

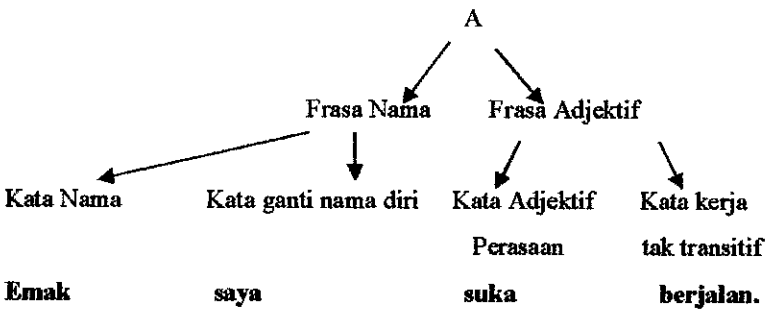
No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak sedang berdiri	Mother is standing.		
2.	Emak sedang berjalan.	Mother is walking.		
3.	Emak sedang menangis.	Mother is crying.		



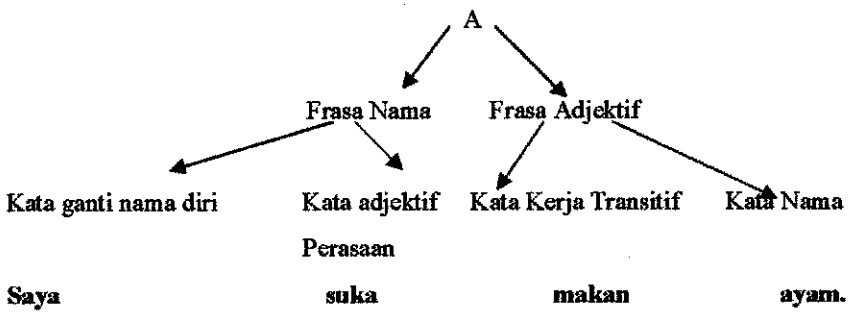
No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Komputer itu sangat mahal.	That computer is very expensive.		
2.	Sungai itu sangat cetek.	That river is very shallow.		
3.	Soalan itu sangat susah.	That question is very difficult.		
4.	Pegawai itu sangat malas.	That officer is very lazy.		
5.	Kuali itu sangat kotor.	That frying-pan is very dirty.		



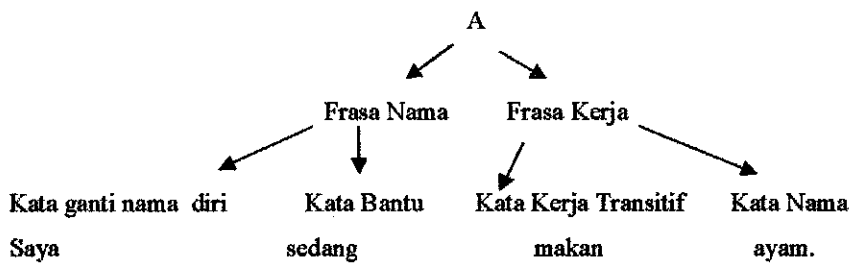
No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak saya sangat gemuk.	My mother is very fat.		
2.	Ayah saya amat pandai.	My father is very clever		
3.	Emak dia sangat kurus.	His mother is very thin		
4.	Ayah mereka sangat hodoh.	Their father is very ugly		
5.	Emak awak sangat busuk.	Your mother is very stinky		



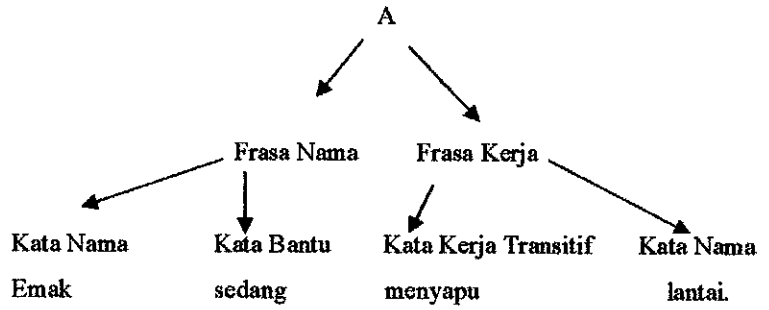
No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak saya suka berjalan.	My mother likes to walk.		
2.	Ayah saya suka berdiri.	My father likes to stand.		
3.	Emak dia suka berjalan.	His mother likes to walk.		
4.	Ayah mereka suka bergurau.	Their father likes to joke.		
5.	Emak awak suka menangis.	Your mother likes to cry.		



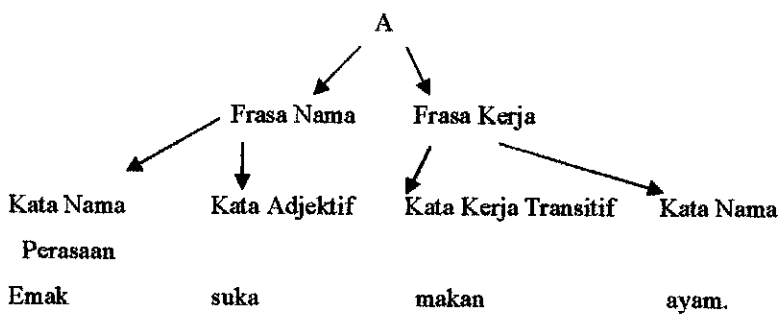
No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Saya suka makan ikan.	I like to eat fish.		
2.	Saya suka membaca buku.	I like to read book.		
3.	Dia suka membaca majalah.	He likes to read magazines.		
4.	Saya suka minum air.	I like to drink water.		
5.	Mereka suka menjahit baju.	They like to sew clothes.		
6.	Mereka suka menendang bola.	They like to kick ball		
7.	Dia suka menyapu sampah.	He likes to sweep rubbish.		
8.	Mereka suka memanjat pokok.	They like to climb tree.		



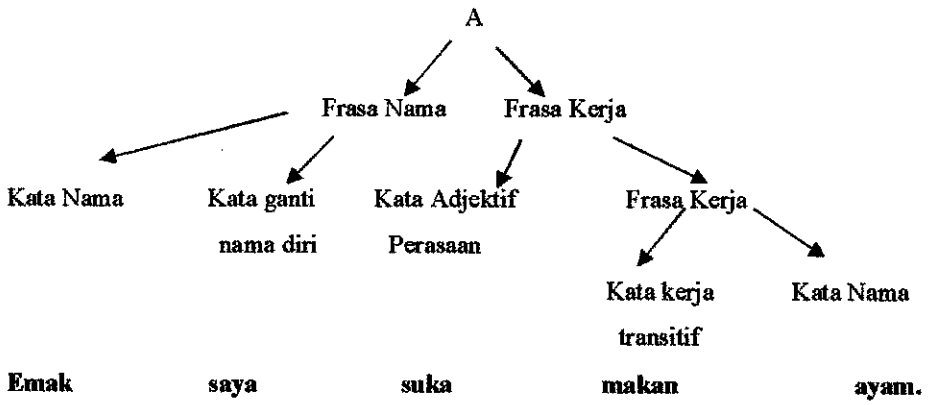
No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Saya sedang makan nasi.	I am eating rice.		
2.	Saya sedang menyapu lantai.	I am sweeping floor.		
3.	Dia sedang memanjat pokok.	He is climbing tree.		
4.	Mereka sedang bermain bola.	They are playing ball.		
5.	Kami sedang mencuci baju.	We are washing clothes.		



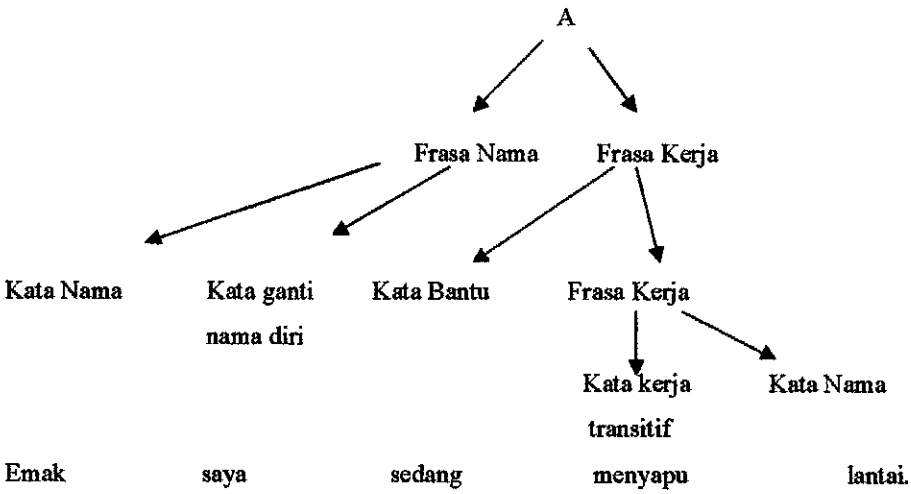
No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak sedang makan nasi.	Mother is eating rice.		
2.	Ayah sedang memanjat pokok.	Father is climbing the tree		
3.	Emak sedang menjahit baju.	Mother is sewing the clothes		
4.	Emak sedang memasak ketam.	Mother is cooking the crab.		
5.	Ayah sedang membuka pintu.	Father is opening the door.		



No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak suka makan sosej.	Mother likes to eat sausage.		
2.	Ayah suka memanjat pokok.	Father likes to climb tree.		
3.	Emak suka membaca majalah.	Mother likes to read magazine		
4.	Emak suka memasak ikan.	Mother likes to cook fish		
5.	Ayah suka menggoreng udang.	Father likes to fry prawn		



No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak saya suka makan ayam.	My mother likes to eat chicken.		
2.	Emak awak suka minum teh	Your mother likes to drink tea.		
3.	Emak saya suka memasak kaserol.	My mother likes to cook casserole		
4.	Emak saya suka menjahit seluar.	My mother likes to sew pants		
5.	Kawan saya suka bermain piano.	My friend likes to play piano		



No.	Malay Sentence (Source)	English Sentence (Target)	Translated grammatically correct.	
			Yes	No
1.	Emak saya sedang makan sosej.	My mother is eating sausage.		
2.	Ayah saya sedang minum kopi.	My father is drinking coffee		
3.	Emak awak sedang menyapu lantai.	Your mother is sweeping the floor.		
4.	Kawan awak sedang memasak sosej.	Your friend is cooking the sausage.		
5.	Abang dia sedang membaca majalah.	His brother is reading the magazine.		

~ Thank you for your time. ~

APPENDIX III: Abbreviation Explanation

Table 7: Abbreviation for Malay Sentence Structure

Abbreviation	Full Name
A	Ayat
FN	Frasa Nama
FK	Frasa Kerja
FA	Frasa Adjektif
KN	Kata Nama
KA	Kata Adjektif
KKT	Kata Kerja Transitif
KKTT	Kata Kerja Tak Transitif
KAPerasaan	Kata Adjektif Perasaan
KGN	Kata Ganti Nama
KGND	Kata Ganti Nama Diri
KHubung	Kata Hubung
KBantu	Kata Bantu
KPenguat	Kata Penguat
KPemeri	Kata Pemeri

Table 8: Abbreviation for English Sentence Structure

Abbreviation	Full Name
S	Sentence
NP	Noun Phrase
VP	Verb Phrase
AdjP	Adjective Phrase
N	Noun
Vt	Transitive Verb
Vint.	Intransitive Verb
A	Adjective
PP	Personal Pronoun
Prep.	Preposition
Aux.	Auxiliary
Det.	Determiner
Int.	Intensifier
C	Conjunction