THE STUDY OF STEMMING ALGORITHM ON MALAY WORDS THAT BEGIN WITH ALPHABETS P, Q, Y AND Z FROM THE TRANSLATED AL-QURAN

SURIANI BINTI MAT

PROJECT SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF BACHELOR OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY
AND QUANTITATIVE SCIENCES
MARA UNIVERSITY TECHNOLOGY
SHAH ALAM

2001

ACKNOWLEDGEMENTS

BISMILLAHIRRAHMANIRRAHIM

In the name of Allah, the Most Gracious and Most Merciful and Him Alone worthy of all praise. Alhamdulillah, thanks you to Allah, this report is finally completed according to time and objectives required.

Here, I would like to take this opportunity to express my most gratitude to those that had helped and inspired me in the completion of this thesis. Special thanks and appreciation is given to my supervisor, for whom I am deeply interested, PM. Dr. Zainab Abu Bakar for her advice, criticism, guidance and creative ideas during the preparation of this project.

To my colleague, thanks for all the moral supports that they had given me and lots many other individuals that unable to be listed who have directly or indirectly help me in completed my thesis.

Last but not least, I would like to express my gratitude to my beloved family for their encouragement, patience, support, financial support and sacrifice they have given me during the course of this thesis.

ABSTRACT

This thesis concerns a Malay language documents retrieval system. Stemming algorithm, Malay Quran translated documents and root dictionaries are used in order to complete this study. The performance of a Malay stemming algorithm is tested based on words beginning with letter 'p', 'q', 'y' and 'z', using 5 experiments. First experiment uses the original set of data collections. In second experiment, new words are added in the dictionary and the total value for 'l', 'm', 'p', 'q', 'y' and 'z' are modified in the header file "dcvarnew.h". Other than that, affixes rule format in file "rule.txt" are added and misspell words are corrected. Third, the locations of rules in file "rule.txt" are changed. For fourth experiment, words that have more than one root, old spelling words and spoken word are deleted from the dictionary. After the modification, the total value for 'k', 'm', 'n' and 'p' in header file "dcvarnew.h" are corrected again. Otherwise, new code is added into module 'ubah_ejaan'. In fifth experiment, the spoken word is deleted from the dictionary and the total value for 'p' in file "dcvarnew.h" is corrected. Then alternative rule to solve the words *pengawal*, *pengawalan* and *perangan* is carried out. The objective of this project is achieved when the best order of the rules to use to stem the words that beginning with p', 'q', 'y' and 'z' is met. This involves the use of two combinations simultaneously such as the pair combination of 1234 as primary combinations and 3124 as the secondary. First, all the words used the 1234 combination, and if the program encountered that the words can not be solved correctly, combination will be shifted to the secondary combination that is 3124 combination. These experiments can serves as a benchmark for future research in Malay language.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Background

A stemming algorithm is a computational procedure, which has the ability to reduce all words with the same root to a common form, usually by discarding each word of its derivational and inflection suffixes (Lovins 1968). Stemming is also carried out in various languages and will be discussed in Chapter II. Popovic & Willett (1992) says that it is one of the well-known conflation algorithms that are used to identify morphological variants.

As for Malay stemming algorithm by Fatimah (1995), it is more effective and powerful than one being developed by Asim (1993). The new stemming approach known as Rules-Application-Order (RAO) approach is the first of its kind to be introduced to cater the stemming process of Malay words by Fatimah (1995).

Asim Othman has the first developed stemming algorithm for Malay words in 1993. By reason of its musical quality, Malay has been styled the "Italian of the orient" (Pei 1968). It has also been described as the world's easier language (Porter 1968). It has no harsh consonant cluster and very few difficulties of grammatical