THE STUDY OF STEMMING ALGORITHM FOR MALAY WORDS THAT START WITH THE LETTER 'B' FROM TRANSLATED QURAN DOCUMENTS

AZRIANA BINTI AHMAD

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

FACULTY OF INFORMATION TECHNOLOGY
AND QUANTITATIVE SCIENCES
UNIVERSITI TEKNOLOGI MARA
SHAH ALAM

1999

ACKNOWLEDGEMENTS

Firstly, I would like to thank Allah for giving me time and strength to finish writing this project.

I would like to express my sincere grateful and gratitude to my supervisor, Professor Madya Dr.Zainab Abu Bakar for her invaluable guidance, encouragement and advice during the course of this project.

My thanks also to all the lecturers that are involved in this project. And not to forget to all the supports by the laboratory staffs for their cooperation in finishing this project.

I would like to thank to all my friend especially Sanisah, Fazlina, Elly Johana, Aniza and Rohana for their cooperation in spending time together to discuss about this project that is under the same area of interest. Also to my housemates at Cermai Street in Section 4 Shah Alam who have given their considerations in finishing this project.

Last but not least, I would like to thank my mother and my father Ahmad bin Zainal Abidin for their encouragement, patience, support and sacrifice they have given me during the course of this project.

ABSTRACT

Stemming algorithm have been developed in many different languages. The Malay stemming approach has been named Rules-Application-Order. The performance of this Malay stemming algorithm is tested using translated Quran documents and Malay dictionary. The words that start with the letter 'b' are extracted from the translated Quran documents as test data. Then, using these words in the experiment on the stemming algorithm is tested employing all the possible combination of affixes rules. The percentages of each combination are compared. The main objective of this study is to test the different combination of affixes for words that start with 'b'. Findings such as words that are wrongly spelt in the translated Quran documents are corrected. Other problems that exist are words that could not be stemmed at all using all the combination. The result of this experiment is to help for future work to correct the problems that are detected for the word that starts with the letter 'b'.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Stemming algorithms are very important because they increase the efficiency of document retrieving systems and reduce the size of index files due to their grouping of many morphological term variants into one single stem. Since a single stem typically corresponds to several full terms, by storing stem instead of terms, compression factors of over 50% can be achieved (Frakes 1992). However, using a stemming algorithm does not always guarantees the improvement of search effectiveness in all circumstances (Harman 1991).

Asim Othman has 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 nature, with no conjunctions or declarations, roots mostly of two syllabus, a consonant-vowel arrangement, concept of gender, number and case generally absent, and loan words mainly from Sanskrit and Arabic. Malay verbs are distinguished by