DECLARATION

I, **Amrita Bhattacharjee**, bearing Registration Number.-Ph.D/2054/12 dated 12/09/2012, hereby declare that the subject matter of the thesis entitled "**Computational Characteristics of Words over Formal Languages**" is the record of works done by me and that the contents of the thesis did not form the basis for award of any other degree to me or to anybody else to the best of my knowledge. The thesis has not been submitted in any other University / Institute. This thesis is being submitted to Assam University for the degree of Doctor of Philosophy in Computer Science.

(AMRITA BHATTACHARJEE)
Research Scholar
(Ph.D. Registration No.-Ph.D/2054/12)

Place:
Date:

Acknowledgements

The endless thanks goes to Almighty for all the blessings He has showered onto me, which has enabled me to write this last note in my present study. I am grateful to Him for giving me the mental and physical strength needed to complete this assignment.

It gives me immense pleasure to express my heartiest sense of gratitude towards my supervisor Dr. Bipul Syam Purkayastha for his invaluable advice, constant encouragement, guidance and support without which this work would not have been possible. From him I have learnt all the necessary basic concepts and ideas for making this Mathematics oriented Computer Science study possible. He has given me the right amount of freedom throughout the present study.

I thank Kh. Raju Singha, Mr. Abhijit Paul, Mr. Arindam Dey, Md. Joynal Abedin and some other fellow scholars for their continuous support.

I also convey my thankfulness to the Head of the Department, all the faculty members and staff in the Department of Computer Science, Assam University, Silchar for their help on various occassions.

I am thankful to several anonymous referees of various journals and conferences for their valuable suggestions and comments which have immensely enhanced the quality of my work.

I thank all my family members, specially to my husband, son and daughter, for their support and encouragement. Finally it's time to thank and pay gratitude to my parents and elders for their blessings, and active cooperation.

— Amrita Bhattacharjee

Dedication

This thesis is dedicated to my beloved

parents

Smti Archana Bhattacharjee and

Shri Lakshmikanta Bhattacharjya,

parents-in-law

Smti Nilima Chaudhuri and Late Nikhiles Chaudhuri,

husband

Biplab Chaudhuri,

son and daughter

Biprajyoti and Adrija

For their endless love, support, encouragement and sacrifice ,

without whom none of my success would be possible.

Contents

Declaration i

Certificate ii

Abstract iii

Acknowledgements vii

Dedication viii

List of Symbols xiii

List of Tables xiv

List of Figures xv

1 Introduction 1

1.1 Motivation: 4

1.2 Statement of the problem: 6

1.3 Objectives: 7

1.4 Data and Methodology: 7

1.5 Preliminary: 8

2 Review of Literature 16

2.1 Basics of Formal language: 17

2.2 On Finite state automata: 20

2.3 On Parikh matrix: 21

2.4 On development of context-free grammar for natural languages: 25

3 Finite State Automaton: A Tool to Represent Formal language 28

3.1 Importance of Finite state automata: 28

3.2 Use of Finite state automata to describe Regular languages: 30

3.3 An Algorithm to construct Deterministic Finite state automata: 34

3.3.1 Result Analysis: 35

3.4 Conclusion of the Chapter: 36

4 On Parikh Matrices of words over formal alphabet 37

4.1 Algorithm to display Parikh matrix corresponding to a Formal word: 39

4.1.1 Algorithm for a Binary sequence: 40

4.1.2 Application of the algorithm for Binary sequence: . . . 41

4.1.3 Algorithm for a Ternary sequence: 42

4.1.4 Application of the algorithm for Ternary sequence: . . 44

4.1.5 Algorithm for a Tertiary sequence: 45

4.1.6 Application of the algorithm for Tertiary sequence: . . 47

4.2 Algorithm to display M-ambiguous words corresponding to a Parikh matrix: 48

4.2.1 Algorithm for a Binary sequence: 48

4.2.2 Application of the algorithm for Binary sequence: . . . 50

4.2.3 Algorithm for a Ternary sequence: 51

4.2.4 Application of the algorithm for Ternary sequence: . . 52

4.3 Introduction of the notion of Graphical representation of Formal words: 53

4.3.1 Introduction of the notion of Two dimensional representation of Binary words: 53

4.3.2 Result Analysis: 55

4.3.3 Introduction of the notion of Three dimensional representation of Ternary words: 57

4.4 Set of equations to find Formal words corresponding to a Parikh matrix: 58

4.4.1 Set of equations to find Binary words corresponding to a Parikh matrix: 58

4.4.2 Set of equations to find Ternary words corresponding to a Parikh matrix: 61

4.4.3 Set of equations to find Tertiary words corresponding to a Parikh matrix: 64

4.5 Definition of Stepping distance on classes of M- ambiguous words: 68

4.5.1 Stepping distance on classes of M-ambiguous Binary words: 68

4.5.2 Stepping distance on classes of M-ambiguous Ternary words: 69

4.5.3 Stepping distance on classes of M-ambiguous Tertiary words : 70

4.6 A few observations on Parikh matrices: 71

4.7 Ratio property and Weak ratio property of sequences: 73

4.7.1 Generalization of Ratio property: 74

4.7.2 Generalization of Weak-ratio property: 75

4.7.3 Searching M-ambiguous words using Ratio property: . 76

4.8 Solving M-ambiguity of words using M-ambiguity Reduction Factor: 80

4.8.1 Representation of M- ambiguous words by Parikh matrix together with M-ambiguity Reduction Factor: . . . 80

4.8.2 Example of M-ambiguity Reduction factor for Binary sequence: 81

4.8.3 Examples of M-ambiguity Reduction factor for Ternary sequence: 83

4.8.4 Example of M-ambiguity Reduction factor for Tertiary sequence: 85

4.9 Conclusion of the Chapter 85

5 Parikh matrix and Natural languages 87

5.1 Representation of Bengali letters by Parikh matrices: 90

5.2 Representation of Bengali words by Parikh matrices: 92

5.3 Representation of Bengali sentence by Parikh matrix: 95

5.4 Conclusion of the Chapter: 96

6 Conclusion 98

Bibliography 102

A List of Publications 112

B List of Implementation of Codings 114

List of Symbols

$\emptyset$	The empty set
$\mathbb{N}$	The set of natural numbers
$\mathbb{Z}$	The set of integers
FSA	Finite state automata
NFSA	Non-deterministic finite state automata
DFSA	Deterministic finite state automata
Σ	An ordered alphabet
Σ^*	The set of words formed from Σ
$\|w\|_{a_i}$	The number of occurrences of a_i in a word $w \in \Sigma^*$
$\Psi_{M_n}(\zeta)$	$n \times n$ Parikh matrix over an word ζ
m_{ij}	The element in the i^{th} row and j^{th} column of a matrix
$\backsim_r$	Ratio property
$\backsim_{wr}$	Weak ratio property
NLP	Natural language processing
$R(\zeta)$	M-ambiguity reduction factor of the word ζ
$d_S(\alpha, \beta)$	Stepping distance between the words α and β
CFG	Context-free grammar
i.e.	That is

List of Tables

3.1 State transition table for an infinite input ‘tuu!’ 31

3.2 State transition table for a finite input hi! 33

3.3 State transition table for a finite input ‘hello!’ 33

List of Figures

1.1 Chomsky hierarchy 11

3.1 Relation of Finite automata, Regular grammars, Regular expressions and Regular languages 29

3.2 Finite state automaton for infinite word ‘tuu!’ 31

3.3 Finite state automaton for the word ‘hi!’ 32

3.4 Finite state automaton for the word ‘hello!’ 33

3.5 Finite state automaton of the word ‘abcd’ 35

3.6 Finite state automaton of the word ‘book’ 35

3.7 Finite state automaton of the word ‘technology’ 36

4.1 Graphical representation of $abbaab$ 54

4.2 Graphical representation of three M-ambiguous words 55

4.3 Graphical representation of four M-ambiguous words 55

4.4 Graphical representation of ten M-ambiguous words 56

4.5 Graphical representation of sixteen M-ambiguous words . . . 56