DECLARATION

I, **Shri Yumnam Bablu Singh**, bearing Registration Number **PHD/1162/2010** dated 15/09/2010, hereby declare that the work presented in this thesis entitled **"WEB ENABLED MULTILINGUAL MANIPURI DICTIONARY"** is the record of work done by me under the guidance of **Prof. Bipul Syam Purkayastha (Supervisor)**, HOD, Department of Computer Science, Assam University, Silchar, Assam and **Prof. Chungkham Yashawanta Singh (Co-Supervisor),** HOD, Linguistics Department, Manipur University, Canchipur, Manipur. I further declare to the best of my knowledge that no part of the thesis has been submitted for the award of any degree to any other University or Institution.

Place: ……………

Date: ……………

(Yumnam Bablu Singh)

Research Scholar

To my parents

Mr. Yumnam Ibobi Singh & Mrs. Y. Ibebi Devi

ACKNOWLEDGEMENTS

In theatrical show, the actors provide their talent to the production and enjoy, as well as suffer, the spotlight. The show could not exist as a vehicle for the actors, however, without the backstage personnel. At the very outset, I am thankful to the authorities of ***Assam University, Silchar, Assam****, for giving me the opportunity to do this dissertation work.*

I am deeply grateful to my guide ***Prof. Bipul Syam Purkayastha****,* ***Computer Science Department****,* ***Assam University, Silchar,*** *for his untiring guidance and valuable advice during the course of my dissertation work. In spite of his various preoccupations, he has kindly spared much of his valuable time to make me more complete and has exposed me to the real world of research work. Without his support, guidance and encouragement my dissertation work would not be in its present form.*

I also thank ***Prof. Hemachandran, Dr. Sahin Ara Begum, Sunita Sarkar*** *and all staff of Department of Computer Science, Assam University, Silchar.*

I am also thankful to my co- supervisor ***Prof. Chungkham Yashwanta Singh****,* ***Linguistic Department, Manipur University, Canchipur*** *for guiding me in my thesis in Linguistic areas in my thesis work and providing data for my thesis from his project on Morphological Analysis of Indian Languages, University of Hyderabad and Indian Language Corpora Initiative (ILCI), J. N. U, Delhi.*

I would like to thank my friends Kh. Raju Singha, Churjit Thiyam, Poireiton S, Shantikumar N., Krishnabati Devi and well-wisher who have helped me directly or indirectly in completion this research work.

I would like to dedicate the successful completion of the thesis work to my parents Yumnam Ibobi Singh and Yumnam Ibebi Devi, two brothers Rishikanta and Misra singh and sister Surbala for supporting me financially, for always encouraging me with positive thought even when I lost my hope as well as and for their unlimited help rendered to me.

Yumnam Bablu Singh

Date……………..

Place……………..

CONTENTS

List of Tables x

List of Figures xi

List of Abbreviations xiii

Abstract xvi

CHAPTER			Page no.
I	**INTRODUCTION ………………………….**		1
		Introduction	1
	1.1	Dictionary	2
	1.2	Multilingual Dictionary	3
	1.3	Rationale of study	4
	1.4	Levels of Structure	6
	1.5	The Challenges in Dictionary Development	7
	1.6	Why Multilingual Dictionary for Manipuri	8
	1.7	Motivation	9
	1.8	Objectives	11
	1.9	Data and Methodology	12
	1.10	Main Contribution	13
	1.11	Thesis Outlines	13
	1.12	Conclusion	14
II	**LITERATURE REVIEW ……………………**		15
		Introduction	15
	2.1	Preliminary Investigation	15
		2.1.1 General Dictionary	17
	2.2	Manipuri Language	19
		2.2.1 The Cultural and Linguistic setting of Manipur	19
		2.2.2 Linguistic Affinity	20
		2.2.3 Manipuri Sounds	21
		2.3.1 Vowels Phonemes of Manipuri	21
		2.3.1.1 Monothongs of Manipuri	22
		2.3.1.2 Dipthongs of Manipuri	22
		2.2.4 Manipuri Script	22
		2.2.5 Cluster	24
		2.2.6 Evolution of Meitei Mayek	24
		2.2.7 Vowels and Consonants	25
	2.3	English and Manipuri Sounds	25
	2.4	English Language	26

2.4.1 Phonology of English 26

2.5 Hindi Language 27

2.6 Dictionaries available in Manipuri Language 28

2.6.1 Graphical representation of Manipuri printed Dictionaries 34

2.7 Hindi Dictionaries 34

2.8 Major English dictionaries 35

2.9 Multilingual Dictionaries 36

2.10 Manipuri Language Technology 37

2.11 Conclusion 37

III **WORDNET AND MMD. ……………………** 39

Introduction 39

3.1 WordNet 39

3.1.1 Concept of lexical Matrix 40

3.1.2 Database Contents 41

3.1.3 Knowledge Structure 43

3.1.4 Psychological Justification 43

3.1.5 WordNet as Ontology 44

3.1.6 WordNet Features 45

3.1.7 Four Kinds of Entailment 47

3.1.7.1 Synonym 47

3.1.7.2 Antonymy 49

3.1.7.3 Hyponymy 50

3.1.7.4 Meronymy 51

3.1.8 Gradation 52

3.1.9 Morphological Relations 53

3.1.10 Synset Syntax 54

3.2 IndoWordNet 55

3.3 Advantages and Limitations of WordNet 56

3.3.1 Advantages of WordNet 56

3.3.2 Limitations of WordNet 57

3.4 WordNet Finding 58

3.4.1 Results of WordNet Data 59

3.4.2 Results of MMD Data 60

3.5 Multilingual Dictionary: MMD 60

3.5.1 MMD Process 61

3.5.2 Lexicon of MMD 62

3.5.3 Process Diagram of MMD with WordNet 62

3.5.4 Advantages of MMD 63

3.4.5 Disadvantages of MMD 65

3.6 Estimate Data for MMD with WordNet 66

	3.7	Conclusion	67
IV	**CORPUS FOR MMD ………………………...**		68
		Introduction	68
	4.1	Corpus	68
		4.1.1 Regular Expression in Corpus	69
	4.2	Motivation for Corpus	71
	4.3	Basic Principles of Corpus Building	73
	4.4	Methodology	74
	4.5	Characteristic Features of MMD Corpus	74
	4.6	Domain Types	76
		4.6.1 Methods of Data Input	78
		4.6.2 Meta Data for Domains	79
	4.7	Types of Corpus	80
	4.8	Corpus Types with regards to user	81
	4.9	Types of Corpus- types of tool	81
	4.10	Parallel Corpora	82
		4.10.1 Advantages of Parallel Corpora	83
	4.11	Corpus Size	84
	4.12	Corpus Schematic Diagram	85
	4.13	CQL (Corpus Query Language)	85
	4.14	Corpus Cleaning/Sanitation	86
		4.14.1 Wild Cards in Corpus	87
		4.14.2 Codes that work in the find what box only (when wildcards are off)	88
	4.15	Schematic Diagram of Corpus Processing	88
	4.16	CQL (Corpus Query Language) Result	89
	4.17	Conversion of raw text to WX- schema	94
		4.17.1 Raw text Corpus	94
		4.17.2 Text converted in WX Schema/notation	94
	4.18	Frequency	94
	4.19	Annotation for POS tagging	96
		4.19.1 Annotation Process	97
		4.19.2 Tagged corpus	97
		4.19.3 CQL Tagged corpus	98
	4.20	Hardware and Software	98
	4.21	Application areas of Corpus	99
	4.22	Conclusion	101
V	**DICTIONARY DESIGN AND SEARCHING**		102
		Introduction	102
	5.1	System Design	102

5.2 Sequential Structure 105
5.3 Trie Structure 106
5.3.1 Trie Search Algorithm 108
5.3.2 The Algorithm of the Trie generation 109
5.3.3 Advantages of Tries over Binary search trees 111
5.3.4 Advantages of Tries over Hash tables 112
5.3.5 Complexity Analysis 112
5.4 Suffix Tree 113
5.4.1 Functionality 114
5.4.2 Search for Strings 114
5.4.3 Find Properties of the Strings 115
5.4.4 Applications 116
5.5 Implementation 116
5.6 MMD Design 116
5.6.1 Page Hierarchy of MMD 117
5.7 Process diagram of MMD 118
5.8 Process diagram of SL and TR language 118
5.9 Tools 119
5.10 Word Entry 121
5.10.1 Algorithm for Word entry 121
5.11 Word class in English, Hindi and Manipuri 122
5.12 Aligning MMD and WordNet 124
5.12.1 Pseudo code of Automatic Alignment Algorithm 124
5.13 Macrostructure and Microstructure of MMD 125
5.14 Output Page of Word entry 127
5.15 Trie Search 127
5.16 Output Page of Search 129
5.17 Statistical calculation for Word entry in MMD 132
5.18 Conclusion 133
VI FUTURE WORK AND CONCLUSION 134
Introduction 134
6.1 Conclusion 134
6.2 Future Work 136
6.2.1 Entry List 136
6.2.2 MMD Corpus 137
BIBLIOGRAPHY 138
APPENDIX A 147
APPENDIX B 148
APPENDIX C 149
APPENDIX D 150

LIST OF TABLES

Table No.	NAME OF TABLE	Page no.
Table 1	**Consonant Phonemes**	27
Table 2	**Hindi Primary, Secondary and Nasal**	28
Table 3	**List of Manipuri Dictionary in Hard Copy**	33
Table 4	**Available Multilingual Dictionaries and its Links**	36
Table 5	**Concept of Lexical Matrix in WordNet**	40
Table 6	**Four Kinds of Entailment in WordNet**	47
Table 7	**Antonymy Table**	50
Table 8	**Meronymy Relationship**	52
Table 9	**Gradation Relationship**	53
Table 10	**WordNet tables in the Database**	54
Table 11	**Number of Words in WordNet**	59
Table 12	**Number of Words in MMD and WordNet**	59
Table 13	**Estimated Data for MMD with WordNet**	66
Table 14	**Types of Corpus user and their needs with regards to the types of Corpus**	81
Table 15	**Wildcards in Corpus Sanitation**	87
Table 16	**Table codes when wildcards are off**	88
Table 17	**Table of Word Frequency Count**	95
Table 18	**Annotation Table**	96
Table 19	**Table of Trie for English Word**	106
Table 20	**Example Trie in English**	107
Table 21	**Example Trie in Manipuri**	108
Table 22	**Word Class in English**	122
Table 23	**Word Class in Hindi**	123
Table 24	**Word Class in Manipuri**	123
Table 25	**Total number of words calculated with WordNet**	132

LIST OF FIGURES

FIGURE No.	NAME OF FIGURE	PAGE
Figure no. 1	Types of Dictionary	3
Figure no. 2	Types of Special Dictionary	16
Figure no. 3	General Dictionary Types	18
Figure no. 4	Schematic Diagram of Sino-Tibetan	20
Figure no. 5	Chart of Monothongs of Manipuri	22
Figure no. 6	Chart of Diphthongs of Manipuri	22
Figure no. 7	Manipuri Script and Characters	23
Figure no. 8	Graphical representation of Manipuri Dictionaries words and Publisher	34
Figure no. 9	Princeton WordNet page for Searching a Lexicon	40
Figure no. 10	WordNet Semantic Relationship in English	46
Figure no. 11	WordNet Semantic Relationship in Manipuri	47
Figure no. 12	Synonymy Diagram	48
Figure no. 13	Antonymy Diagram	49
Figure no. 14	Hyponymy Diagram	51
Figure no. 15	Meronymy diagram	52
Figure no. 16	Gradation Level	52
Figure no. 17(a)	IndoWordNet Home page	55
Figure no. 17(b)	Search page of IndoWordNet	56
Figure no. 18	Graphical representation of WordNet Data	59
Figure no. 19	Graphical representation of English, Hindi and Manipuri Data	60
Figure no. 20	Stages for building Multilingual Dictionary	61
Figure no. 21	Step by step process for MMD development	62
Figure no. 22	Lexicon of Manipuri words	62
Figure no. 23(a)	Schematic diagram how WordNet data are used for MMD	63
Figure no. 23(b)	Process diagram of WordNet to MMD	63
Figure no. 24	Graphical representation of WordNet Data	66
Figure no. 25	FSA Diagram of Corpus	69
Figure no. 26	Workflow of parallel corpus of MMD	84
Figure no. 27	Schematic Diagram of Corpus Processing	85
Figure no. 28	Process Diagram of Web based Corpus	86
Figure no. 29	Schematic diagram of Corpus processing	89
Figure no. 30	Output Page of CQL	89
Figure no. 31	Frequency of Word in CQL	90
Figure no. 32	Word Occurrence in CQL	90
Figure no. 33	Frequency Graph of Corpus Processing	91
Figure no. 34	Concordance of word in Corpus processing	91
Figure no. 35	Frequency distribution of Corpus processing	92
Figure no. 36	Word list of Corpus processing	92
Figure no. 37	Word concordance of Corpus processing	93
Figure no. 38	Occurrence of particular word	93
Figure no. 39	Annotation process for POS tagging	97

Figure no. 40	CQL for POS in Manipuri Corpus	98
Figure no. 41	Words formation with root word and Affixes	105
Figure no. 42	A Trie Structure of English words with reference to table no. 19	107
Figure no. 43	A trie structure of Manipuri words	108
Figure no. 44	Context diagram for www.mmd.com	117
Figure no. 45	Hierarchical page chart of MMD	117
Figure no. 46	Diagram showing raw data collection for MMD database	118
Figure no. 47	Parallel translation of Source language to Target language	119
Figure no. 48	Word entry category of MMD	121
Figure no. 49	Word entry page	127
Figure no. 50	Output page search by English word	129
Figure no. 51	Search by Manipuri word	129
Figure no. 52	Search by English word for WSD	130
Figure no. 53	Search result by Manipur for WSD	130
Figure no. 54	Search result by English to Manipuri	131
Figure no. 55	Search result by Manipuri	131
Figure no. 56	Graph showing maximum no. of word entry for MMD	132
Figure no. 57	MMD the multilingual Corpus: English-Hindi-Manipur	137

LIST OF ABBREVIATIONS

ABL	Ablative
ACC	Accumulative
AIATSIS	Australian Institute of Aboriginal and Torres Strait Islander Studies
API	Application Programming Interfaces
ASCII	American Standard Code for Information Interchange
ASP	Aspect
ASP.NET	Active Server Page dot net
AMD	Anglo Manipuri Dictionary
BE	Be Verb
BN	Benefactive
BNC	British National Corpus
BIS	Bureau of Indian Standards
CAB	Case Ablative
CAUS	Causative
CFILT	Center for Indian Language Technology
CIIIL	Central Institute of India languages
CL	Computational Linguistics
CN	Case
CQL	Corpus Query Language
CRF	Conditional Random Field
CON	Concession
COP	Copula
CC_CCD	Conjunction_Co-ordinator
CC_CCS	Conjunction_Subordinator
CC_CCS_UT	Conjunction_Quotative
C-DAC	Centre for Development of Advanced Computing
DA	Deitic
DFA	Deterministic Finite Automata
DIST	Distal
DIT	Department of Information Technology

DM_DMD	Demonstrative Deictic
DM_DMR	Demonstrative Relative
DM_DMQ	Demonstrative_Wh-word
ETL	Electronic Text Library
EMPH	Emphatic
EVI	Evidential
FEMD	Friends English Manipuri Dictionary
GNU	General Public License
GUI	Graphical User Interface
IE	Information Extraction
ILCI	Indian Languages Corpora Initiatives
INTRA	Interrogative
IR	Information Retrieval
IPA	International Phonetic Alphabet
JJ	Adjective
OALD	Oxford Advanced Learner's Dictionary of Current English
OCR	Optical Character Recognition
PDE	Determiner
POS	Part of Speech
PPF	Post Position
PSP	Postposition
RB	Adverb
RD_SYM	Residuals Symbol
RD_PUNC	Residuals Punctuation
RD_UNK	Residuals Unknown
RD_RDF	Residuals_ForeignWord
RP_RPD	Particles Default
RP_CL	Particles Classifier
RP_INJ	Particles Interjection
RP_INTF	Particles Intensifier
RP	Received Pronunciation
RP_NEG	Particles Negation
SAMD	Student Anglo Manipuri Dictionary

SGML	Standard Generalized Markup Language
SL	Source Language
SQL	Structured Query language
TTF	Thin Film Transistor
TG	Together
TL	Target Language
LSI	Linguistics Survey of India
LU	Language Unit
MEPLA	Manipuri Equivalents of Parliamentary, legal and administrative Terms
MESAT	Manipuri Equivalents of Scientific and Technical Terms
MSD	Morphosyntactic Descriptions
MMD	Multilingual Manipuri Dictionary
MRD	Machine Readable Dictionary
MT	Machine Translation
MWE	Multi Word Expression
NEG	Negation
NFA	Non- Deterministic Finite Automata
NLP	Natural Language Processing
NMZ	Nominalizer
OED	Oxford English Dictionary
SOV	Subject Object Verb
SVO	Subject Verb Object
TB	Tibeto-Burman
URL	Uniform Resource Locator
WSD	Word Sense Disambiguation
WX	Indian Language in ASCII format
V_CV	Cunjunc verb
V_CV_C	Compound verb
V_CV_SPC	Conjunc verb with spa