PERMANENT REFERENCE

- Language Assistant -

Towards Automatically Translating Text from English to Tamil in a Controlled Environment

LIBRARY
2 6 MAR 1999
Eastern University, Sri Lanka.

001.4 072
MOH

By

S . Mohanarajah

(93 / M.Sc.Com.Sc / 32)

ExMSc23
Thesis
Library - EUSL

31405

Dissertation Submitted in
Partial Fulfilment of the Requirement for the

M.Sc in Computer Science

at the

Department of Statistics and Computer Science
University of Colombo, Sri Lanka

June – **1998**

PROCESSED
Main Lib

PERMANENT REFERENCE

Abstract

Automatic translation is the process of automatically converting a text in one natural language to another using computers. The resultant translated text should tally with the original text in meaning, effect and style. Since this automated translation process involves handling a huge amount of dictionary information, and complex grammar and word formation rules, the vocabulary and sentence structures of the input text may be restricted based on domain specific knowledge to get a better translation.

Sri Lanka is a multi-ethnic country. Communication between the ethnic groups is essential for co-operation in the interest of national development. Translation is the only viable solution for communication between communities which use different languages. Sri Lanka requires a six-way translation tool between Tamil, English and Sinhala. The major objective of this study, as a first step for the overall motivation, is to build a model for a Machine Translation system that could translate text in English to Tamil in a controlled environment.

The proposed system is built on transformer based architecture. A code for information interchange in Tamil (SCIIT) is designed and used in this project. The system developed in this project works well for the English sentence structures found in elementary level textbooks. This system may be used readily for demonstration purposes, however, it needs further enhancements for it to be used in a real translation task.

Contents

Abstract

Acknowledgment

Declaration

Contents

CHAPTER I

PROJECT OVERVIEW 1

Introduction 2

Motivation........ 4

Scope of Study........ 5

Methodology........ 5

Deliverables 6

Dissertation Overview........

CHAPTER II

MACHINE TRANSLATION BASICS 8

Introduction 8

Human Languages - Some Basic Features........ 10

Language as a Symbolic System. *11*

Natural Language Processing (NLP)........ *12*

Machine Translation. 13

Linguistic Knowledge........ 14

Phonology: 15

Morphology:........ 16

Syntax 16

Phrase Structure Analysis........ 18

Parsing : 18

Grammatical Relations Structure Analysis: 20

Semantic 22

Pragmatic, Discourse and Real world knowledge: 24

Generation (Synthesizing) *24*

Translation Engines........ 25

Transformer Engines. (Direct Systems)........ 26

Transfer Engines........ 27

Interlingua Engines. *29*

Dictionaries and MT systems........

CHAPTER III

LINGUISTIC ANALYSIS ON ENGLISH

Introduction. 30

English Phonemes, Syllables and Stresses. 30

English Morphemes and Word formations. 32

English Word Classes. 33

English Phrases, Clauses and Sentences. 34

Nouns and Noun phrases. 35

Pronouns. 37

Adjective and Adjective Phrases. 38

Adverbs and Adverb Phrases: 39

Prepositions and Prepositional Phrases: 39

Verb and Verb Groups 40

Negation 42

Modal Verbs 42

English Sentence Structure. 43

Case I : Subject + Verb 43

Case II: Subject + Verb + Complement 43

Case III: Subject + Verb+ Object 44

Case IV: Subject + Verb+ Object+ Ascriptive Complement 44

Case V : Subject + Verb+ DirectObject + IndirectObject 44

CHAPTER IV

LINGUISTIC ANALYSIS ON TAMIL

Introduction 45

Tamil Language Fundamentals 45

Tamil Alphabets, Phonemes, and Syllables 46

Roots of Tamil Words 49

Word Combination Rules in Tamil 51

Tamil Word Classes 53

Tamil Nouns 54

Caste 54

Gender 54

Number 55

Cases in Tamil Nouns 55

Pronouns and Cases 57

Tamil Verbs 58

Transitive, Intransitive and Causal Verbs 60

Passive Voice in Tamil 61

Negation in Tamil 61

.....62
Modal Verbs.....62
Adjective Particles.....63
Adverbial Particles. .....

CHAPTER V

DESIGN ISSUES *64*
Introduction *64*
Interface Design 64
User Characteristics.....65
Levels of Interfacing.....66
Form Link Diagram (FLD). *66*
Functional Design..... *67*
SCIIT - Simple Code for Information Interchange in Tamil..... *69*
English Morphological Process. *71*
Tamil Morphological Process. 74
Gender Codes..... *75*
Dictionary.....77
Dictionary Look-up (English to Tamil & Tamil to English) *79*
Translation. 80
Input Related Issues.....81
Controlled Environment. 81
Tockenization..... *83*
Translating Engine. 83
English Language Parser. 84
Verb Groubs 86
Grammatical Relations Structure. 89
Verb Code..... *91*
Comparative Grammar Rules. *96*
Sentence and Document translator utilities. .....

CHAPTER VI

RESULTS AND DISCUSSION *98*
Introduction *98*
1 . *Dictionary Look-Up Utilities*.....99
1.1 English to Tamil Dictionary.....100
1.2 Tamil to English Dictionary..... *101*
2. *Translation Utilities.* 102
2.1 Sentence Translator 109
2.2 Document Translator 113
Translation Mechanism .....

CHAPTER VII

LIMITATIONS AND FUTURE EXTENSIONS

Introduction *114*

Limitations. *114*

Limitations on Implemented software 115

Designed, but not Implemented portions 119

Other Basic Requirements for an English to Tamil MT system 119

Future Extensions *120*

CHAPTER VIII

SUMMARY & CONCLUSION

Introduction *124*

Summary *124*

Conclusion. *127*

Bibliography........ i

Appendix A : Program Code........ iv

Appendix B : Interface Design........ lix

Appendix C : English Phonemes........ lxv

Appendix D : Tamil Phonemes........ lxvi

Appendix E : SCIIT - Simple Code for Information Interchange in Tamil........ lxviii

Appendix F : Tamil Keyboard Structure - Sarawathy........ lxix

Appendix G : Some Tamil Word Formation Rules........ lxx

Appendix H : Noun Cases (Tamil) - Overview........ lxxiii

Appendix I : Synopsis of Pronouns........ lxxv

Appendix J : Tamil verb Conjugates........ lxxvii