

**USING MULTI AGENT TECHNOLOGY FOR
AUTOMATIC MACHINE TRANSLATION**

Budditha Hettige

(118036M)

Degree of Doctor of Philosophy

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

July 2020

USING MULTI AGENT TECHNOLOGY FOR AUTOMATIC MACHINE TRANSLATION

Budditha Hettige

(118036M)

Thesis submitted in partial fulfillment of the requirements for the degree
Doctor of Philosophy

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

July 2020

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or part in print, electronic or another medium. I retain the right to use this content in whole or part in future works (such as articles or books)



13.07.2020

.....

.....

Signature:

Date:

Budditha Hettige

Candidate

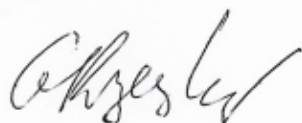
The above candidate has carried out the research for the PhD thesis under my supervision.

.....

.....

Prof. Asoka S. Karunananda

Date:



14.07.2020

.....
Prof. George Rzevski

.....
Date:

Dedicated to

This thesis is dedicated

...to my beloved mother and father

...to my wife and son

Acknowledgement

Many people have helped their best to successfully completion of this research. I acknowledge all of them for their valuable thoughts, and constant encouragement gave me to make my research a reality.

First and foremost, I acknowledge my supervisor senior professor Asoka Karunananda for accepting me as his research student and giving excellent support and advice. Prof. Karunananda is a great mentor who guided while giving me all the freedom and encouragement to accompany with my ideas. Also, I acknowledge my second supervisor, professor George Rzevski for accepting me as his research student and giving excellent support and advice. Without both them patient listening and creative thoughts, this work would not have been possible at all.

Especially, I acknowledge Venerable Kirioruwe Dhammananda Thero for his kind-hearted help and encouragement to fulfil my research work and give correct direction to my successful life.

My very special thank goes to Dr (Mrs.) Uditha Rathnayake, Dr (Mrs.) Menaka Ranasinghe and Dr Lochandraka Ranathunga for their invaluable comments and guidance as my examiners of bi-annual review panels.

Also, I wish to extend my sincere thanks for the support I received from all the members of the administration office and members of the Faculty of Information Technologies, University of Moratuwa. Especially I thank Dr (Mrs.) Thushari Silva and Dr (Mrs.) Subha Fernando and Madam, Prof. Deelika Dias for their essential roles.

I also acknowledge Mr P. Dias (Department of Statistics, University of Sri Jayewardenepura) and Ms. E.R.C. Sadamali for their kind support to fulfil my research work.

Also, exceptional and heartfelt thanks for Dr. (Mrs.) Mihirini Wagaarachchi, Dr (Mrs.) Chinthani Weerakoon, Dr. (Mrs.) Anuradha Ariyarthne and Ms. Mihiri Serisooriya for their gracious associations throughout the last couple of years.

I would like to thankfully remind all the academic and non-academic members of the General Sir John Kotelawala Defence University those who have supported me during the work. Special thanks go to the Vice-chancellor Major General Milinda Peiris, the Dean, Faculty of Computing, Commodore J. U. Gunaseela, for granting me the study leave to complete my research.

Again, I would like to mention Dr Asele Gunasekara, Maj R.M. Rathnayaka and all the members of my faculty gave me a great support

Finally, I would like to extend my greatest gratitude to my family members, especially my wife Lakshmi and my little hart Tenuja for the unrestricted support given, and without their care, this would have been unmanageable. Again, I must give an express thanks to Lakshmi for tolerating my busy schedules due to the research work. Last but not least, I thank all who supported me to make this work a success.

July 13, 2020

Budditha Hettige

Abstract

Machine translation is a cost-effective, quick, and widely accepted automated language translation method that has become essential in the modern and ever more globalized world. Machine translation can be done with one or more different approaches, including dictionary-based, rule-based, example-based, phrase-based, statistical, or neural-linguistic approaches. Nevertheless, most of the existing machine translation systems show a quality gap when compared with human translation. Thus, human translation has been considered as the best language translation method so far. Human language translation is a complex and opportunistic process depends on human memory. This human language translation process has been described through a few theories. Among them, the garden path model and the constraint satisfaction model are two fundamental approaches available for human language translation, especially concerning sentence parsing with meaning. These two theoretical models demonstrate how to select suitable words in the phrase of a sentence to generate accepted meanings. Based on these two theories, a hybrid approach to machine translation has been proposed. This proposed approach is stimulated by how people parse and translate a sentence by putting available phrases together with accepted meaning. According to the approach, translation is done in three stages. In the first stage, the system analyses the given sentence by considering the morphology, syntax, and semantics of the source language. Then, the system uses phrase-based translation and translates each phrase into the target with multiple solutions. The phrase translation is done considering the four factors of psycholinguistic parsing techniques, such as phrase structure, semantic features, thematic roles, and probability. Finally, considering all the translated phrases, the system should be capable of identifying suitable target language phrases to take accepted meanings, considering subject-verb and object-verb agreements. After the subject-verb-object agreement, other available phrases in the sentence should be capable of re-arranging according to the accepted subject, object, and verb phrases.

This approach has been simulated with the multi-agent system named EnSiMaS, which translates English text into Sinhala. The EnSiMaS was implemented on the MaSMT framework, which was specially developed for agent-based machine translation. The EnSiMaS comprises of 26 language processing agents on both source and target languages. These agents were clustered into six agent swarms considering morphological, syntactical, and semantical concerns of the source and the target languages. In addition to these language-processing agents, the system should be able to create an agent dynamically for each source language phrase. These dynamically created phrase agents should be capable of communicating with other relevant phrases and taking the accepted solutions.

The EnSiMaS was tested with 85 sample English sentences. For each English sentence, three different translations were taken. According to the evaluation result, the system shows an 8.77% word error rate, a 6.72% inflexion error rate, and a 5.37% sentence error rate for the first, second, and third translations. In addition, calculated BLUE scores show 0.89160756, 0.52009204, and 0.43581893 for the first, second, and third translations. Then randomly selected 25 samples sentences are used to calculate the adequacy and fluency of the EnSiMaS. Adequacy and fluency rates were taken from 55 human evaluators considering the human-translated reference sentences. The Kendal's Tau correlation coefficient shows that there is a weak positive association between adequacy levels of human translations vs EnSiMaS system translations and moderate positive association between fluency levels of human translation and EnSiMaS system translation. Further, according to the Fleiss Kappa coefficient method, there is a significant fair agreement on raters for adequacy and fluency ratings.

Keywords: Machine Translation, Multi-agent systems, Human Language Processing, MaSMT, EnSiMaS

Table of Contents

Declaration	i
Acknowledgement	iii
Abstract	v
Table of Contents	vi
List of Figures	xiii
List of Tables	xvi
List of Abbreviations	xviii
CHAPTER 1	1
INTRODUCTION	1
1.1 Prolegomena	1
1.2 Aim and Objectives	3
1.3 Problem in Brief	4
1.4 The Scope of the Research	4
1.5 Hypothesis	4
1.6 Human Translation to Machine Translation	5
1.7 Proposed Approach to Machine Translation	6
1.8 Resource Requirements	7
1.9 Chapter Organisation	8
1.10 Summary	9
CHAPTER 2	10
STATE OF THE ART IN MACHINE TRANSLATION	10
2.1 Introduction	10
2.2 Fundamentals of Machine Translation	10
2.3 Brief History	12
2.4 Existing Approaches to Machine Translation	14
2.4.1 Interlingua approach	14
2.4.2 Human-Assisted/Computer-Aided Translation	16
2.4.3 Dictionary-based Machine Translation	18
2.4.4 Rule-based Machine Translation	19
2.4.5 Example-based Machine Translation	21

2.4.6 Statistical Approach to Machine Translation	22
2.4.7 Neural Machine Translation	24
2.4.8 Knowledge-based Approach	27
2.4.9 Transfer-based Machine Translation	27
2.4.10 Agent-based Approach to Machine Translation	28
2.4.11 Hybrid Approach to Machine Translation	29
2.5 Local Resource and Existing ESMTS	30
2.6 Some Issues in Machine Translation	31
2.6.1 Word and Sentence Segmentation	31
2.6.2 Word Conjugation	31
2.6.3 Tense Detection	32
2.6.4 Multi-word Expression	32
2.6.5 Out of Vocabulary	32
2.6.6 Translating Idiomatic Phrases	32
2.7 Summarization of Existing MT Approaches	33
2.8 Problem Definition	34
2.9 Summary	34
CHAPTER 3	35
LITERATURE REVIEW AND BACKGROUND	35
3.1 Introduction	35
3.2 Computational Grammar for the English Language	35
3.2.1 The Morphology of the English Language	35
3.2.2 Syntax of the English Language	40
3.2.7 The Semantics of English Language	42
3.3 The Sinhala Language	44
3.3.1 Morphology of the Sinhala Language	44
3.3.2 Syntax of the Sinhala Language	48
3.4 Comparison Between English and Sinhala Languages	49
3.5 Summary	49

CHAPTER 4	50
NATURAL LANGUAGE PROCESSING TECHNIQUES	50
4.1 Introduction	50
4.2 Computational Model for English and Sinhala	50
4.3 Morphological Analysis and Generation	53
4.4 Syntactical Analysis and Generation	55
4.5 Semantics Processing	56
4.5.1 Word level semantics	56
4.5.2 Phrase level semantics	57
4.5.3 Sentence level semantics	57
4.6 Summary	57
CHAPTER 5	58
MULTI AGENT TECHNOLOGY	58
5.1 Introduction	58
5.2 What is Multi-agent System?	58
5.2.1 Type of Agents	59
5.2.2 Agent Communication	60
5.3 Existing MAS Development Framework	60
5.4 MaSMT: Multi-agent Framework for Machine Translation	63
5.4.1 MaSMT Framework	64
5.4.2 AGR Organisational Model and MaSMT Architecture	64
5.4.3 MaSMTAbstractAgent	66
5.3.4 MaSMT Agent	66
5.3.5 MaSMT Agent’s Life cycle	67
5.3.6 MaSMT Controller agent	67
5.3.7 MaSMT Root Agent	67
5.3.8 MaSMT Agents’ Swarm	68
5.3.9 MaSMT Messages	69
5.3.10 MaSMT Settings	70
5.3.11 MaSMT Message Parsing	71
5.3.12 Applications of MaSMT	72

5.4 Summary	72
CHAPTER 6	73
A HYBRID APPROACH TO MACHINE TRANSLATION	73
6.1 Introduction	73
6.2 A Novel Approach to Machine Translation	73
6.3 Theoretical Basis of Language Translation	74
6.4 A multi-agent Approach to Machine Translation	76
6.4.1 Multi-agent Approach to English Morphological Analysis	76
6.4.2 Multi-agent Approach to English Syntax Analysis	76
6.4.3 Multi-agent Approach to English to Sinhala Phrase-based Translation	77
6.4.4 Multi-agent Approach to Sinhala Morphological Generation	77
6.4.5 Multi-agent Approach to Sinhala Syntax Generation	78
6.5 Why a Multi-agent Approach?	78
6.6 Features of EnSiMaS	79
6.7 Input for EnSiMaS	79
6.8 Output of EnSiMaS	79
6.9 Process of the EnSiMaS	80
6.10 Summary	81
CHAPTER 7	82
DESIGN OF THE ENSIMAS	82
7.1 Introduction	82
7.2 Design of the EnSiMaS	82
7.2.1 EnSiMaS GUI	83
7.2.2 Ontology	83
7.2.3 Virtual World	84
7.2.4 English Morphological Swarm	85
7.2.5 English Syntax Analysing Swarm	86
7.2.6 Bilingual Semantics Swarm	87
7.2.7 Sinhala Morphological Swarm	88
7.2.8 Sinhala Syntactical Swarm	89
7.2.9 Ontological Swarm	89

7.2.10 English Phrase-based Translation Swarm	90
7.3 Summary	91
CHAPTER 8	92
IMPLEMENTATION OF ENSIMAS	92
8.1 Introduction	92
8.2 EnSiMaS Ontology	92
8.2.1 English Pronoun Table	93
8.2.2 English Regular Noun table	94
8.2.3 English Regular Verb Table	94
8.2.4 English Irregular noun table	95
8.2.5 English Irregular verb table	95
8.2.6 English Regular Adjective Table	96
8.2.7 Other word table	96
8.2.8 English-Sinhala Bilingual Dictionary	97
8.2.9 Morphological rules for English words	97
8.2.10 Syntax Rule for English phrases	98
8.3 EnSiMaS Virtual World	99
8.3.1 The English Word and Word List	100
The English Word Morphology	101
8.3.2 The English Phrase and Phrase List	101
8.3.3 The Sinhala Phrase and Sinhala phrase list	102
8.3.4 The Sinhala word Lexicon	103
8.3.5 The EnSiMaS Phrase	104
8.3.6 The EnSiMaS Phrase List	105
8.3.7 The EnSiMaS Sentence info	106
8.4 EnSiMaS Agents	106
8.4.1 EnSiMaS Manager Agent	106
8.4.2 English Morphological System	107
8.4.3 English Syntax Swarm	110
8.4.4 Bilingual Semantic Swarm	113
8.4.5 Sinhala Morphological Generation Swarm	115

8.4.6 Sinhala Syntactical Swarm	117
8.4.7 Translation Controller Agent	117
8.4.8 Translation Swarm	118
8.4.9 Ontological Swarm	119
8.5 Summary	120
CHAPTER 9	121
EVALUATION	121
9.1 Introduction	121
9.2 English to Sinhala Multi-agent System (EnSiMaS)	121
9.3 EnSiMaS Dictionary	122
9.4 EnSiMaS Translator	123
9.5 EnSiMaS Phrase-based Editor	128
9.6 Evaluation Strategy of the EnSiMaS	129
9.6.1 Round Trip Translation	130
9.6.2 Word Error Rate	130
9.6.3 Sentence Error Rate	131
9.6.5 Inflectional Error Rate	132
9.6.6 BLEU	132
9.6.7 METEOR	133
9.6.8 Human Evaluation	133
9.7 Experiment	137
9.8 EnSiMaS vs Google Translator	140
9.9 Results and Data Analysis	141
9.9.1 Details of the sample set	141
9.9.2 Adequacy and Fluency	143
9.10 Conclusion of the Data Analysis	149
9.11 Summary	150
CHAPTER 10	151
CONCLUSION AND FURTHER WORK	151
10.1 Introduction	151
10.2 Hybrid Approach for Machine Translation	151

10.3 Conclusion	152
10.4 Objectives-wise Achievement	154
10.5 Limitations	157
10.6 Further Work	157
10.7 Summary	158
References	159
Appendix A: Translation Summary with Agents' Communications	180
Appendix B: EnSiMaS User Manual	198
Appendix C: Sample of evaluation form	201
Appendix D: List of Publications	204
Appendix E: MaSMT Development Guide	205

List of Figures

Figure 2.1: General pipeline of the MT	10
Figure 2.2: Machine translation pyramid	11
Figure 2.3: Historical Development of the MT	13
Figure 2.4: Taxonomy of the Machine Translation	14
Figure 2.5: Translation Process of the interlingua MT	15
Figure 2.6: General pipeline of the dictionary-based MT system	18
Figure 2.7: Design of the dictionary-based MT	19
Figure 2.8: Components of the RBMT system	20
Figure 2.9: Activities on statistical MT	23
Figure 2.10: Encoder-decoder architecture of the NMT	25
Figure 3.1: Part of speech mapping between English and Sinhala	44
Figure 4.1: Language model for English and Sinhala	51
Figure 4.2: Ontology of a word	51
Figure 4.3: Ontology for a Phrase	52
Figure 4.4: Ontology for a Sentence	53
Figure 4.5: Process of the morphological analysis and generation	55
Figure 5.1: Different types of agents	59
Figure 5.2: UML-Based Aalaadin model for multi-agent system development	64
Figure 5.3: Agents' architecture on MaSMT	65
Figure 5.4: Modular architecture of the MaSMT Agent	66
Figure 5.5: The life cycle of the MaSMT Agent	67
Figure 5.6: Architecture of the MaSMT controller agent	68
Figure 5.7: Design of the swarm of Agents	69
Figure 6.1: Factors contribute to sentence parsing	74
Figure 7.1: Design of the EnSiMaS	82
Figure 7.2: Design of the EnSiMaS Ontology	84
Figure 7.3: Design of the EnSiMaS Virtual world	85
Figure 7.4: Design of the English morphological swarm	86
Figure 7.5: Syntax analyzing swarm: agents for English syntax analysis	87
Figure 7.6: Design of the Bilingual Semantics swarm	88

Figure 7.7: Design of the Sinhala morphological swarm	89
Figure 7.8: Design of the phrase-based translation Swarm	90
Figure 8.1: Structure and sample data on the pronoun table	93
Figure 8.2: Structure and sample data on the regular noun table	94
Figure 8.3: Structure and sample data on the regular verb	94
Figure 8.4: Structure and sample data on the irregular noun	95
Figure 8.5: Structure and sample data on the irregular Verb table	95
Figure 8.6: Structure and sample data on the regular adjective table	96
Figure 8.7: Structure and sample data on the other word table	96
Figure 8.8: Structure and sample data on the bilingual dictionary	97
Figure 8.9: sample data for English morphological rules	98
Figure 8.10: Selected rules to detect English phrases (Phrase rules)	99
Figure 8.11: Class diagrams of the English word and English wordlist	100
Figure 8.12: Class diagrams of the English word morphology	101
Figure 8.13: Class diagrams of the English phrase and English phrase list	102
Figure 8.14: Class diagrams of the Sinhala Phrase and Sinhala phrase list	103
Figure 8.15: Class diagrams of the Sinhala word lexicon and Sinhala word lexicon list of the EnSiMaS	104
Figure 8.16: Class diagram of the EnSiMaS phrase	105
Figure 8.17: Class diagram of the EnSiMaS phrase list	106
Figure 8.18: Activity diagram of the morphological agent	108
Figure 8.19: Communication diagram of the EMS	109
Figure 8.20: Activity diagram of the English Syntax Swarm	112
Figure 8.21: Communication diagram of the syntactical swarm	113
Figure 8.22: Communication diagram of the Bilingual semantics swarm	114
Figure 8.23: Activities of the Sinhala Noun Generation agent	116
Figure 8.24: Agent communication diagram of the translation swarm	119
Figure 9.1: Top-level application selection GUI of the EnSiMaS	121
Figure 9.2: GUI of the EnSiMaS dictionary	122
Figure 9.3: A dictionary-based bilingual word editor	123
Figure 9.4: GUI of the EnSiMaS Translator	128
Figure 9.5: GUI of the EnSiMaS phrase-based editor	129

Figure 9.6: Distribution of the number of words among input sentences	142
Figure 9.7: Percentage distribution on adequacy and fluency values for EnSiMaS Best Translation	145
Figure 9.8: Distribution on adequacy rates on five different raters	146
Figure 9.9: Distribution on fluency rates on five different raters	147

List of Tables

Table 2.1: Summary of the selected MT systems	33
Table 3.1: Some Inflectional suffixes in English	36
Table 3.2 Some Morphological rules for Noun Inflection	37
Table 3.3 Regular and irregular noun forms	37
Table 3.4: Regular and irregular English verb forms	38
Table 3.5: Some Morphological rules for Verb conjugation	38
Table 3.6: Adjective relationship of a noun	39
Table 3.7: Verb and adverb usage	39
Table 3.8: Basic Thematic Relationship in a sentence	43
Table 3.9: Sinhala Noun inflexion form for base word මුඛ (dear)	45
Table 3.10: Verb inflexion forms for Verb <i>maranawa</i> (මරණවා)	46
Table 3.11: Add-remove values for the Sinhala verb	47
Table 3.12: Fundamental differences in both Sinhala and English	49
Table 4.1: Summary of the Existing Morphological analyzers	54
Table 5.1: summary of the existing Multi-agent system development frameworks	63
Table 5.2: Structure of the MaSMT Messages	70
Table 5.3: Default settings of the MaSMT	70
Table 5.4: Message directives (headers for messages)	71
Table 8.1: Statistics of the EnSiMaS Knowledgebase	92
Table 8.2: Agents' details of the English morphological swarm	107
Table 8.3: Morphological Tags	109
Table 8.4: Agents' details of the English syntax swarm	110
Table 8.5: Agents' details of the Bilingual Semantic swarm	114
Table 8.6: Agents' details of the Sinhala morphological swarm	115
Table 8.7: Sinhala Syntax Generation Swarm	117
Table 8.8: Ontological Swarm	120
Table 9.1: 1-5 Scale Adequacy matrix	134
Table 9.2: Fluency value in the Likert scale	134
Table 9.3: Fleiss' Kappa values for agreements	135

Table 9.4: Fleiss' Kappa values for agreements	136
Table 9.5: 25 Sample sentences with translated results	138
Table 9.6: Comparison between EnSiMaS vs Google Translator	139
Table 9.7: Summary of descriptive statistics of the 85 input sentences	141
Table 9.8: Calculated WER, IER and SER for the translations	142
Table 9.9: Calculated BLEU results for each translation	143
Table 9.10: Fleiss' kappa coefficient values for Adequacy	144
Table 9.11: Fleiss' kappa coefficient values for Fluency	145
Table 9.12: Summary of the Kendall's rank correlation coefficient for adequacy between Human translation and EnSiMaS translation	148
Table 9.13: Summary of the Kendall's rank correlation coefficient for fluency between Human translation and Fluency on EnSiMaS Translation	149

List of Abbreviations

AI	- Artificial Intelligence
ARGM	- Agent Role Group Model
BCE	- Before the Current Era
BEES	- Bilingual Expert for English to Sinhala
CE	- Current Era
CAT	- Computer Assisted Translation
CYK	- Cocke–Younger–Kasami
CSM	- Constraint Satisfaction Model
EMA	- English Morphological Analysis
ESA	- English Syntax Analysis
ESMTS	- English to Sinhala Machine Translation System
EnSiMaS	- English to Sinhala Multi-Agent System
EBMT	- Example Based Machine Translation
FIPA	- Foundation for Intelligent Physical Agents
GNMT	- Google’s Neural Machine Translation
GPM	- Garden Path Model
HAMT	- Human-assisted (-aided) machine translation
IER	- Inflexion Error Rate
JADE	- Java Agent DEvelopment Framework
KQML	- Knowledge Query and Manipulation Language
LSTM	- Long Short Term Memory
LL	- Left-to-right, Leftmost derivation
LR	- Left-to-right, Rightmost derivation
MWE	- Multi-Word Expressions
MAS	- Multi-Agent System
MT	- Machine Translation
MaSMT	- Multiagent System for Machine Translation
NMT	- Neural Machine Translation
NLTK	- Natural Language Toolkit
NPMT	- Neural Phrase-based Machine Translation

NLP	- Natural Language Processing
PPO	- Preposition Phrase Order
RBMT	- Rule-based Machine Translation System
RTT	- Round-trip Translation
SL	- Source Language
SMT	- Statistical Machine Translation
SER	- Sentence Error Rate
SMG	- Sinhala Morphological Generation
SSG	- Sinhala Syntax Generation
SOV	- Subject Object Verb
SVO	- Subject Verb Object
SPADE	- Smart Python multi-Agent Development Environment
TAG	- Tree Adjoining Grammar
TL	- Target Language
WER	- Word Error Rate

CHAPTER 1

INTRODUCTION

1.1 Prolegomena

Language translation is a process of communication of meaning from one language into another using a human or a machine [1]. The word translation is derived from the Latin word translation, which refers to the meaning of carrying or bringing across [2]. Thus, translation can be defined as carrying or bringing meaning from one language to another. Language translation tasks have more than two thousand years of a long history. Historically, in the Asian region, a Buddhist monk, Kumārajīva [3], translated Buddhist texts written in Sanskrit to Chinese in 868 Common Era, which was the first-ever recorded translation task in the Asian languages. In the Western world, a collection of Jewish Scriptures translated into early Koine Greek in Alexandria has been considered as the first recorded attempt for language translation [4].

In the early days, all the translation tasks were conducted by humans who had sound knowledge of both source and target languages [5]. However, considering the limited human resources and time taken for the translations, human-based language translation takes more cost [6]. Further, current statistics of the world show that there are listed 7,111 living languages and 3,995 languages have writing systems [7]. Thus, language translation requirements of the present globalised world cannot be sufficient through only the human translation [8]. It requires accurate machine translation systems. As a result of this, automated machine translation is more popular than human-based language translation [9].

Machine Translation (MT) is an automated process of the language-translation which was done by the Machines. MT is classified under the task of Natural Language Processing (NLP) in the area of Artificial Intelligence (AI)[10]. Further, In general, MT systems translate one language text (source language) into another language (target language) considering the meaning of the source language text. MT systems consider the given input source and identify morphological, syntactic, and semantic relations. According to the morphological, syntactical, and semantic information

available on the source language sentence(s), machine translation systems are capable of identifying a suitable target language context [11].

Further, number of well-known approaches are available for machine translation, including dictionary-based [12], rule-based [13], phrase-based [14], interlingua [15], Knowledge-based [16], Example-based [17], Statistical [18], Neural-MT [19], and Hybrid [20]. Most of these approaches use different translation techniques and demonstrate different performance for language translation.

Direct transfer, syntax transfer, and interlingua are the three main classifications of the rule-based Machine Translators. Most of the direct transfer systems are responsible only to the morphological level [21]. Therefore, the direct transfer system succeeds only in closely related languages [22]. The syntax level transfer systems are responsible for both morphology and syntax levels [23]. However, less concern goes to the semantic level. Interlingua translation gives attention fully to all these morphological, syntactical, and semantic levels [24].

Further, the rule-based translators follow a set of morphologic, syntactic, and semantic rules to deliver their translation [25]. In general, rule-based systems capable to provide grammatically correct translations. The phrase-based approach shows the morphological and syntactical (especially phrase-level) analysis for the machine translation. The corpus-based approach uses some corpus-based methods (statistical or example-based) for target language selection [26]. The neurolinguistics approach [27] is modern and one of the most successful approaches to MT that follows machine learning techniques. However, it should require more training and more language resources (more resources for both source and target languages) to provide accurate translations.

Further, natural languages are more complex, and they differ from language to language. Thus automated language translation is a challenging task from the last seven decades. In some complicated situations, there is a quality gap between human translation and machine translation [28] [29].

Theoretically and technically, humans' language translation process differs from existing machine translations. The human language translation procedure is considered

as a complex and opportunistic process that is based on human ontology (human memory) [30]. For instance, different people give different translations for the same text. The humans' language parsing is also based on phrase structure, semantic features, thematic roles, and probability [31]. Also, several psycholinguistic models are available to demonstrate the task of language translation. The Garden Path Model (GPM) [32] and the Constraint Satisfaction Model (CSM) [33] are the two models available for human language translation. These models demonstrate how to select suitable words in the phrase of a sentence to take accepted meaning [34].

Note that the GPM is a serial modular parsing model that takes a solution according to the selected meaning at the beginning of the reading. The CSM uses all the available probabilistic information in the sentence at once to take a solution.

Considering GPM and CSM psycholinguistic models, a novel approach is proposed for machine translation [35]. This proposed approach also differs from existing systematic computer-based fully automated machine translation approaches.

This thesis presents a novel approach to machine translation that models human language translation concepts for automated machine translation. The approach has been simulated through the multi-agent system(MAS), EnSiMaS (English to Sinhala Multi-agent System). The EnSiMaS has been implemented through a specially developed multi-agent development framework named MaSMT [36].

This chapter contains a brief overview of the entire thesis including a brief introduction to the field of interest; the problem has been addressed, the aim and objectives, a brief introduction of the proposed approach and layout of the thesis are also presented.

1.2 Aim and Objectives

This thesis proposes a hybrid approach to MT that implements the human language translation approach to machine translation through multi-agent technology. Therefore, the aims of this research to design and develop an English to Sinhala

machine translation system by using Multi-agent System Technology. To achieve the above aim, the following objectives have been recognised:

1. Critically review of existing machine translation approaches and systems.
2. An in-depth study to model Sinhala and English languages to build an ontology for agents.
3. Critically review multi-agent technology for MT.
4. Define a language translation method that is capable of translating English text into Sinhala like a human.
5. Design and develop a machine translation system using multi-agent system technology.
6. Evaluate the system.

1.3 Problem in Brief

Compared with a human translated result, the current machine translation systems produce less quality translation [37]. Thus, there is a translation quality gap between the human translated results and the machine-translated results.

1.4 The Scope of the Research

The scope of this research is limited to develop a MAS that translates English to Sinhala only.

1.5 Hypothesis

Multi-agent technology can be used to design a machine translation system capable of processing morphology, syntax and semantics interactively, like humans, rather than sequentially, as current machine translation systems

1.6 Human Translation to Machine Translation

Language translation is not an easy and straightforward task because of the complexity of natural languages. The humans' language translation requires several types of linguistics knowledge [38] including common-sense knowledge, morphological knowledge, phonological knowledge, pragmatic knowledge, semantics knowledge, and syntactic knowledge [39]. This linguistics knowledge differs from person to person. Note that human language translation is an opportunistic process done by the human to convert meaning from one language into a known another language considering the semantics, syntax, and morphology as required [40]. Theoretically, human translation can be considered with three assertions, namely:

1. Opportunistic rather than algorithmic: The translation process varies from context to context (sentence to sentence). The procedure of the translation is complex, and the algorithmic process cannot apply for the translation process.
2. Based on human knowledge: The result of the translation varies from person to person. For instance, the same sentence can be translated differently.
3. Vary from sentence to sentence: Semantics concern has been done through the sentence level, and the translation is based on the phrase-level [41].

The translation proceeds by the humans, reading word by word left to right (English has a left-to-right word order). By reading words, humans are capable of identifying available phrases and doing some translations. Then by connecting each phrase, the final translation can be obtained.

Note that the phrase-level translation has been considered as one of the best translation methods that have been identified by many researchers [42] [43]. However, according to the complication of the natural languages and the psycholinguistic activities done by humans, the procedure of the translation is not defined yet. In addition, the weight of the above four factors affected for the translation, namely phrase structure,

semantics, thematic rules, and probability, which are involved in the translation, is also unknown. Thus, the identification of the human language translation process is also a research-challenging task.

With the above facts, this research has been conducted through the three stages. As the first stage, the procedure of an English to Sinhala human translation process was identified with human support. Then, a suitable agent model was designed to simulate the humans' translation procedure. Finally, MAS has been developed to simulate the proposed approach.

1.7 Proposed Approach to Machine Translation

The proposed approach is based on how humans translate the given English sentence into Sinhala through psycholinguistic parsing techniques. According to the approach, translation has been done through the hybrid method, including the phrase-based psycholinguistic approach and the multi-agent approach. The phrase-based psycholinguistic approach is powered through the two main psycholinguistic theories for language parsing, namely the garden path model (GPM)[44] [45] and the constraint satisfaction model (CSM) [46].

According to the GPM, The reader takes word by word from left-to-right and generates (Re-build) the meaning according to the existing context using some assumptions while they are reading. Sometimes new information presents on the latter part of the sentence, which is not matched with the existing assumptions [47]. Then it should be required to find another suitable assumption for the existing context. According to the above facts, the GPM is restricted to a single context. However, the CSM uses all the available information at once.

The proposed approach is based on these two models and uses combined methods from the above two models. According to the approach, the translation system takes all the possible solutions for each English phrase available in the input English sentence. Then phrase-level translation has been done considering phrase structure, thematic relation, and probability.

Among translated Sinhala phrases, the best Sinhala translation for the subject phrase is selected considering the probability. After that, a suitable Sinhala verb (through the verb phrase) is selected according to the selected Sinhala subject phrase (use subject-verb communication). The object has been selected according to the previously selected subject and verb. Then rest of the part is selected according to the selected subject, verb, and object.

Multi-agent technology has been used to simulate the above translation process. Through the morphological and syntax processing of the source language, available phrases are identified considering their thematic relations and phrase structure. This source language analysis can be done through the agents, which are capable of identifying morphological and syntactical relationships on the input sentence. Then relevant Sinhala translations are taken for each English phrase with considering the phrase structure, thematic rules for the phrase, and the probability of the usage. The Sinhala morphological generation system also gives support to Sinhala phrase generation. Further, phrase agents are created by putting English phrases and the translated Sinhala phrases into the agent's knowledgebase. These phrase agents are capable of communicating with required other agents to identify suitable Sinhala translation phrases from an existing list of translations. For instance, the subject agent communicates with verb agents and identifies suitable Sinhala verb phrases for the existing subject phrases. Subsequently, the object agent communicates with a relevant verb phrase agent and selects a suitable Sinhala object phrase. After selecting suitable Sinhala phrases, the system should be capable of generating a Sinhala sentence by rearranging the order of the existing translated phrases with support from the Sinhala syntax generation system. Applying the same procedure, the system should be capable of generating multiple Sinhala translations for the different Sinhala subject phrases.

1.8 Resource Requirements

The following software and hardware resources list needed to accomplish the research.

JDK 1.8 or above; SQLite and NetBeans 8.0 have been used to implement the proposed MAS. Further, the entire system has been developed through Java. Therefore, to execute the system, JDK 1.8 or above and SQLite are required.

1.9 Chapter Organisation

The rest of this thesis is arranged as follows.

The second chapter of the thesis reports a literature review on MT with a detailed description of the history of machine translation, existing approaches, systems, and, finally, some issues with machine translation.

In the third chapter of the thesis presents literature of the English and Sinhala languages with a computational model for both languages as per morphological, syntactical, and semantical concerns on both languages.

The fourth chapter of the thesis discusses natural language processing techniques related to machine translation, including morphological analysis, syntax analysis, morphological generation, and syntax generation etc.

The fifth chapter of the thesis discusses the multi-agent system technology, including agents, agent communications, existing systems, and agent development frameworks. Finally, it gives a complete note on the MaSMT framework, which was specially designed for the ESMT.

The sixth chapter reports the proposed approach. This chapter also presents the hypothesis of the project including the input, translation process, output, and the important features of the proposed system.

The seventh chapter is about the design of the proposed EnSiMaS.

The eighth chapter reports the implementation details of the EnSiMaS including English-Sinhala knowledgebase (lexical database), the phrase-based translation tool, and the EnSiMaS classical translator.

The ninth chapter of the thesis explains applications of EnSiMaS”, namely, the English to Sinhala bilingual dictionary, the phrase-based tool, and the EnSiMaS translator. Besides, this chapter also reports an evaluation of the EnSiMaS, including methodology, steps, and the result of the evaluation.

The conclusions and further works are given in chapter ten.

1.10 Summary

This chapter briefly discussed the introduction of the research by highlighting human translation and machine translation. The research problem addressed in the thesis is given along with objectives to the approach proposed by the thesis. At the last section, a brief description of the thesis layout was reported. The next chapter critically reviews existing approaches and systems in machine translation.

CHAPTER 2

STATE OF THE ART IN MACHINE TRANSLATION

2.1 Introduction

The previous chapter presented an introduction to the research by highlighting human translation and machine translation. The research problem addressed in the thesis was also presented along with objectives. This chapter gives a critical review of machine translation, including existing approaches, systems and issues.

2.2 Fundamentals of Machine Translation

Machine translation (MT) is a way of converting the meaning of one language into others through a software program. In general, most machine translation systems are required to analyse the source and generate target while considering the meaning of the source. Thus, machine translation can be demonstrated as a meaning translation procedure, which was done by computers. Figure 2.1 shows the general pipeline of MT.

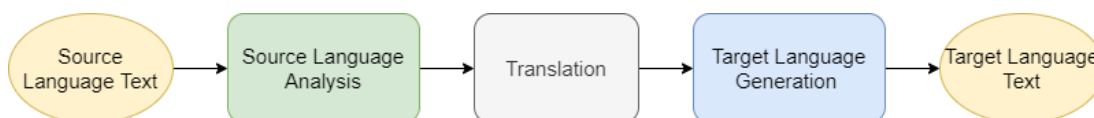


Figure 2.1: General pipeline of MT.

According to this pipeline, the translation process can be considered as a sequential process of the source language analysis into the target language generation. Thus, the translation accuracy may vary from approach to approach and levels of language analysis and generation have been considered in the translation. In this point of view, the process of the machine translation can be categorised into six levels, namely lexical, morphological, syntax, semantics, pragmatic, and interlingua [48]. A lexical level machine translation system only considers lexical entities (most of these translators are dictionary-based translation systems). The morphological level system considers lexical and morphology for both source and target languages. Subsequently,

syntax, semantics, and pragmatic level transfer systems use language processing up to that level. Among all these types of systems, the interlingua system gives complete analysis and generation on all aspects of text processing. Figure 2.2 shows the said diagram of the translation, which is already known as the machine translation pyramid. Source: Interlingua in Google Translate [49].

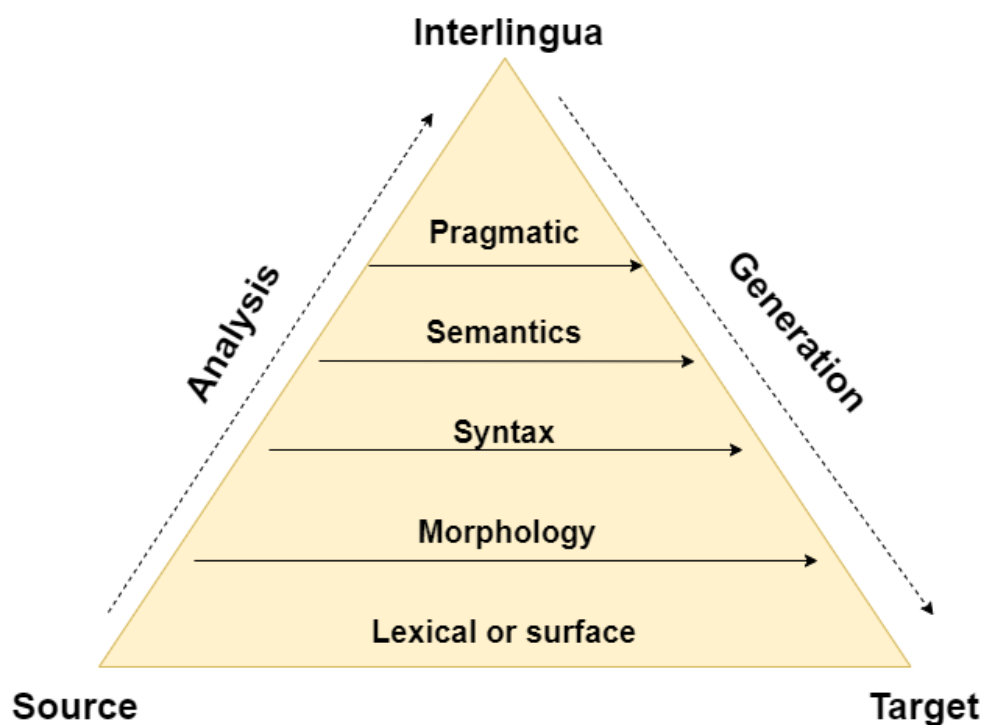


Figure 2.2: Machine Translation Pyramid

Further, several well-known approaches are available for Machine Translation including rule-based, syntax-based, statistical, example-based, and neural linguistics. These approaches take different techniques for language processing. Thus, the quality of the translation may vary from approach to approach.

At the earliest age, rule-based systems are commonly used for machine translation. In the middle age, most of the systems use corpus-based methods with statistical techniques. Neural linguistics (including machine learning capabilities) is the modern and most accurate approach for MT. With this point, a brief description of the history of machine translation has been noted below.

2.3 Brief History

Natural language processing research (especially in MT research) has more than a seventy-year history. From a historical point of view, the dictionary look-up system was the first NLP application that was developed at Birkbeck College, London in 1948 [50]. In the meantime, Brand wood and Cleave made a mechanical calculator [51], which was capable of handling some linguistic problems [52].

In 1948, Booth and Richens introduce a dictionary lookup procedure to handle machine translations [50] using a mechanical dictionary. In 1951, a summary was taken considering issues and benefits of machine translation to satisfy translation demands, particularly in science [53]. According to this summary report, various options for machine translation, including different approaches (mixed MT) and the process of human translation were reported. Besides, this summary report proposed a post-editor, which was considered to handle semantic ambiguities and pointed out an ability to solve grammatical ambiguities automatically through considering the operational syntax[54]. This summary report is the first attempt to take a machine translation process as a human translation. However, existing technologies were not very powerful to simulate the human translation process in those days.

The first machine translation conference was held in 1952 at MIT [55]. At the conference, all participants agreed on the need for pre-editing and post-editing. Also, they agree to give every possible version of the translation and idiomatic phrases [56] should either be included as units in the lexicon, as a solution for the machine translation.

A word-for-word machine translation system for Russian text into English was introduced by the Perry at MIT in 1952 [57] as a simple dictionary lookup system. In the next decade, most researches used a trial-and-error approach to develop machine translation systems [58] for English to other languages. Historically, the first MT system (Russian into English) was developed in 1950 [59]. In 1958, the first practical MT system (Russian text into English) was implemented by the IBM to US Airforce

under the direction of Gilbert King [48]. After 1970, SYSTRAN [60] implemented a new Russian-English MT system to replace the previous MT system of the US Airforce.

In 1980, computer-aided translations were the most successful approach for MT, especially for Japanese- English [61]. After 1980, several machine translation research was done in many areas. Among others, corpus-based machine translation approach has been the most popular approach until now. At present, neural machine translation is the most successful approach for MT which was first introduced by Google in 2016 [37]. Figure 2.3 shows some historical development of machine translation.

The above timeline only shows some selected remarkable developments. The next section reports some popular approaches to machine translation.

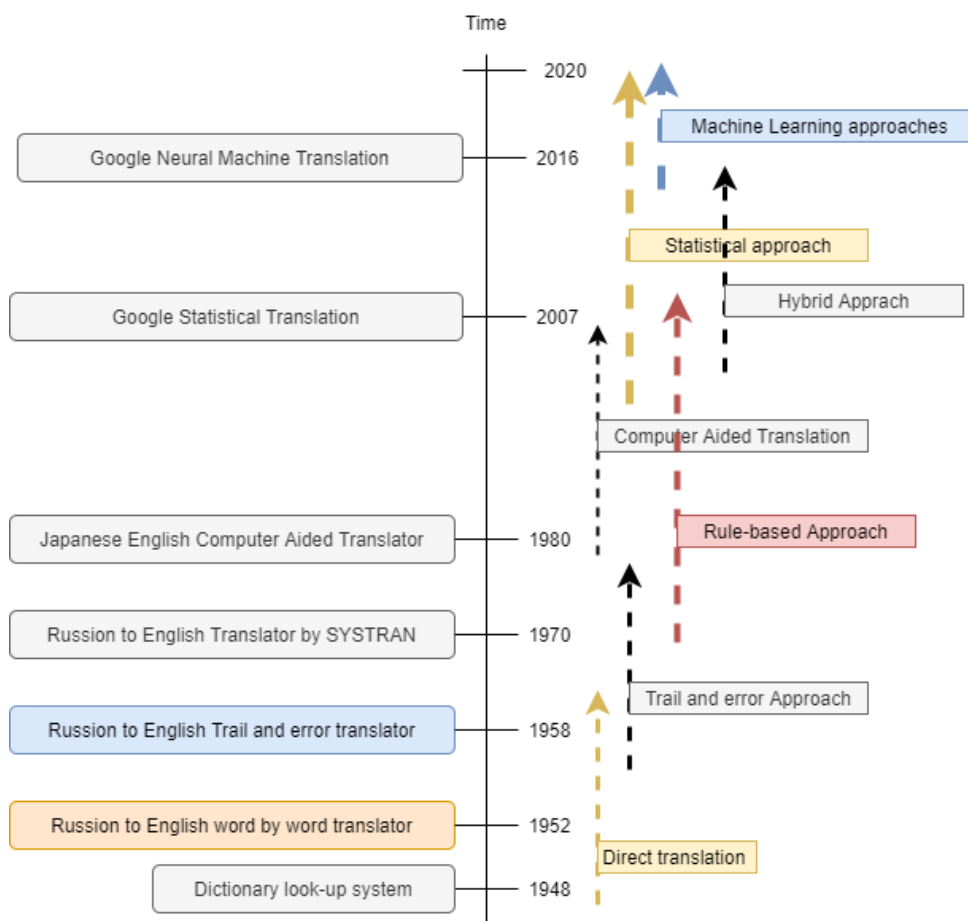


Figure 2.3: Historical Development of the MT

2.4 Existing Approaches to Machine Translation

MT systems can be considered according to their translation approach, or level of language processing that has been done. In general, interlingua, rule-based, human-assisted, and statistical approaches are some common approach to MT. All these MT approaches have their features and limitations. Figure 2.4 shows the taxonomy of machine translation.

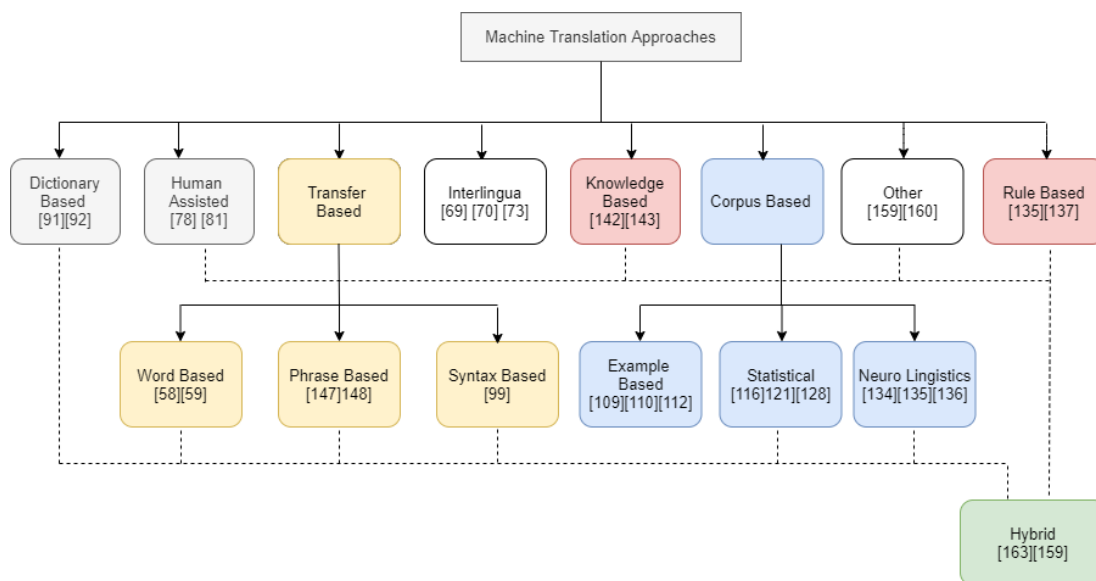


Figure 2.4: Taxonomy of the Machine Translation

2.4.1 Interlingua approach

The machine translation systems can be classified into three main groups, such as direct, transfer, and interlingua[62]. The interlingua approach first takes source language text and translated into abstract language-independent representation called interlingua. Then generates target language form that representation [63]. Generating a language-independent representation for the source text is difficult tasks in practice [64]. However, this approach gives an easier way to add a new language (easily used for multiple language translation) than all other methods. The main limitation of the interlingua approach is the meaning representation of the source language. Note that, if the meaning of the source language is more complex, then generation will be too difficult. The theoretical point of view, the interlingua approach is the top-level

approach of the level of machine translation that is available in the machine translation pyramid. Therefore, an interlingua system requires all the levels of language analysis and generation tasks including morphology, syntax, semantics, and pragmatic levels. Figure 2.5 shows the translation process of the interlingua machine translation.

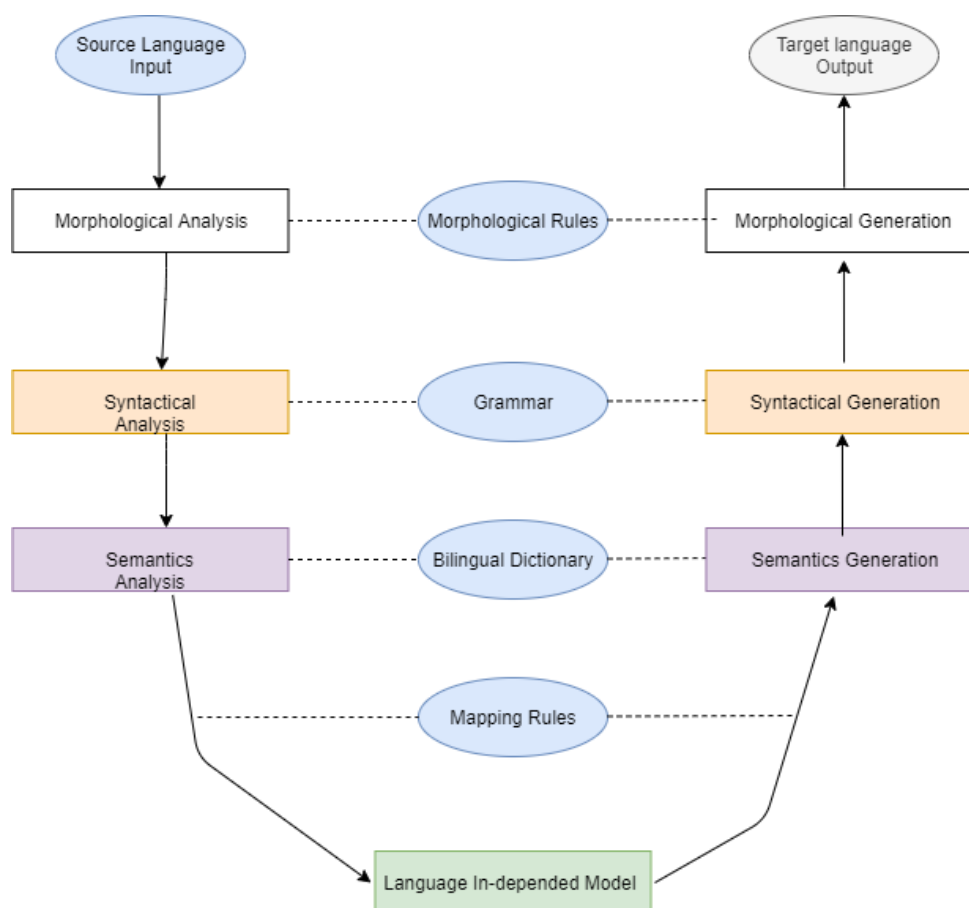


Figure 2.5: Translation process of the interlingua MT

For instance, the numbers of early systems use the interlingua approach for its language translation including English to Arabic MT [65] and Chinese-English MT (ICENT) [66]. Further, The Thai to English MT is another interlingua based machine translation system.[67]. This system translates Thai sentences into interlingua using LFG grammar [68].

In the early stages, UNITRAN [69] was developed as an interlingua system that translates English, Spanish, and German using the parameter-setting method. This

system also characterises language distinctions on both syntactic and lexical-semantic levels.

In the Indian region, the English-Hindi interlingua-based machine translation system was developed [70]. Language knowledge was implemented using the universal networking language [71]. This system should be capable of handling a single sentence at once and defines a “semantic net-like structure” to represent the semantics [72] on the sentence.

An interlingua based MT has been developed for Sanskrit to English [73] using Paninian framework [74]. This system uses Sanskrit text as an input and builds an interlingua. Then it uses interlingua representation for convert (mapping).

Further, machine translation with multiple approaches is a common trend in the field of machine translation. Yichao et al. have successfully applied a neural interlingua technique for multilingual machine translation [15].

Note that, the interlingua approach for machine translation is a bit difficult, which is required a complete knowledge of both source and target languages.

2.4.2 Human-Assisted/Computer-Aided Translation

Computer-assisted or computer aid translation system [75] uses human support for the machine translation, through the pre-editing, post-editing, or intermediate level. These types of systems still used only for low resource language translations. Note that interactive translation where humans and machines co-operate is more successful to archive human-quality machine-translated solutions[76].

The human-assisted approach is much popular for low resource languages, and that gives human interaction (editing) for pre, post or intermediate stages. Therefore, these types of translation systems are considered as semi-automated machine translation systems. There are a number of CAT tools available, including OmegaT [77] and MemoQ. Note that, OmegaT is a free, open-source Java-based Computer-Assisted translation tool (translation memory application) use by the many people to edit and

translate their document easily. The most useful feature was available on the OmegaT including customizable segmentation feature, and spell checking facilities [78].

MemoQ [79] is other popular computer-assisted translation software, that can be run on Windows. MemoQ [80] also provides translation memory, terminology, machine translation integration and reference information management facilities.

Few CAT tools have been developed for Indian languages including, Anglabharti, Mantra and Anusaaraka [81].

Anglabharti [82] is one of the popular machine aided translation system for English to Indian languages. This system uses a pattern directed approach with context-free grammar like structures [83]. This CAT tool provides a human-engineered post-editing package for the target language to make the final corrections. With human support, this approach provides more quality than the transfer approach but less than the interlingua approach.

MANTRA [84] is another Machine assisted translation tool to translate Indian language documents, especially for the selected domain. MANTRA uses Tree Adjoining Grammar (TAG) [85] and Mildly Context-Sensitive Grammar [86] for parsing and generation. In addition to that, MANTRA provides multiple output selection, online word addition, grammar creation, and updating facilities [87].”

At the early stage, Anusaaraka [88] MT provides human aided translation for a few Indian languages with supporting the Paninian Grammar model [89]. Besides, English-Hindi Anusaaraka translates English text into Hindi [90].

Note that, human-assisted translation tools or CAT tools can be considered as a semi-automated tool that provides pre-editing, post-editing, or intermediate editing facilities. Generally, these tools available only for low resources and or complex languages.

2.4.3 Dictionary-based Machine Translation

The dictionary-based MT is one of the early approaches to machine translation categorised under direct translation. The translation process of the approach is based on the dictionaries (resources availability on dictionaries). These translation systems give attention for word level (some systems should be capable of handling morphology); however, not much concern on other levels. In general, the dictionary-based translation is based on word-by-word (word level) translations. For that, this approach is more accurate on languages that are closely related (like Pali and Sinhala languages [91], Czech and Russian [92]). Also, the performance of the dictionary-based translation can be enhanced by introducing the source language morphological analyser and target language morphological generator for the translation. Figure 2.6 shows the general pipeline of the dictionary-based MT system. Further, the improved design of the dictionary-based MT system is shown in Figure 2.7.

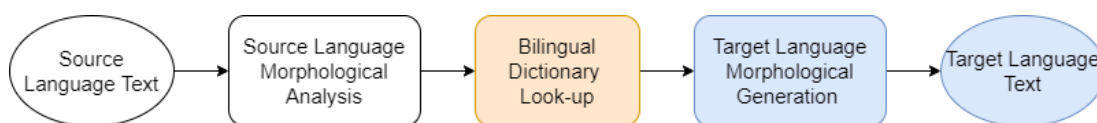


Figure 2.6: General pipeline of the dictionary-based MT

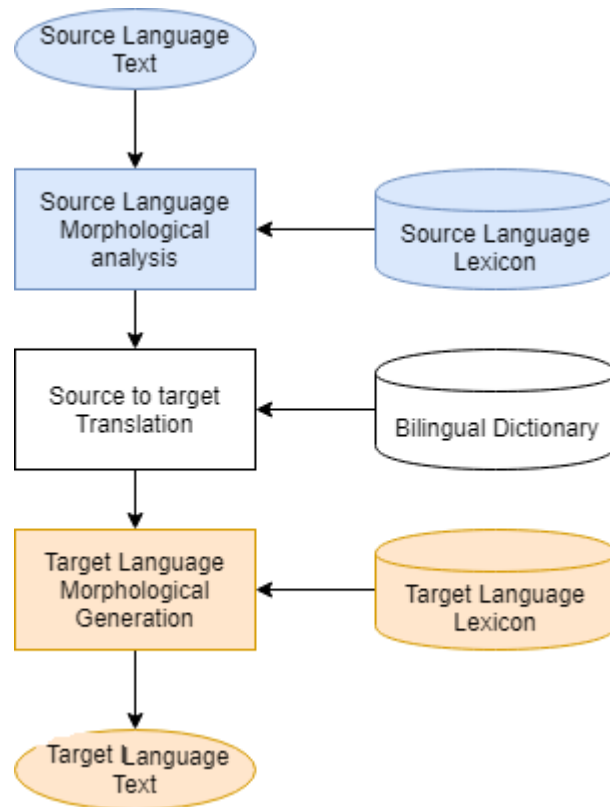


Figure 2.7: Design of the dictionary-based MT

According to the improved design, three dictionaries (source, target and bilingual lexical resources) required to complete the translation. Pali to Sinhala MT can be considered as an application for the dictionary-based MT [91].

2.4.4 Rule-based Machine Translation

The Rule-based MT (RBMT) is the classical approach for MT based on linguistic information about the source and target languages. More importantly, the RBMT system uses a set of language-specific rules to provide grammatically correct translations [93]. In general, the RBMT system contains five main Language processing modules with several lexicon dictionaries [13]. Figure 2.8 shows the general design architecture of the RBMT.

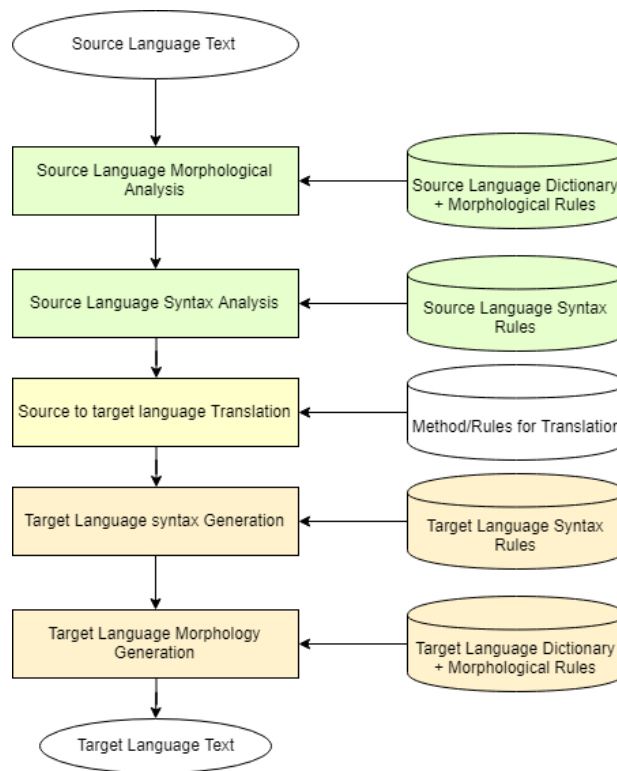


Figure 2.8: Components of the RBMT system

Numbers of MT systems use a rule-based approach. Among others, Apertium [94] [95] is an open-source RBMTS, that translates related languages through the shallow-transfer approach [96]. At present, Apertium commonly used as a toolbox to develop RBMT systems.

Toshiba [97] is another RBMT for English to Japanese vice versa. The Toshiba uses seven steps language processing method with semantic transfer schema to develop grammar system called Lexical Transition Network Grammar [98].

BEES (Bilingual expert for English to Sinhala) is one of the rule-based English to Sinhala Machine Translation System (ESMTS), that can translate grammatically correct English sentence into Sinhala through the direct transfer techniques [99]. The BEES system consists of five language processing modules for machine translation namely, English Morphological analyzer [100], English Parser, English to Sinhala translator, Sinhala Morphological generator and Sinhala composer [101]. In addition, the system uses Sinhala word conjugation rules for Sinhala grammatically correct

word generation [102]. Further, the BEES system uses three Prolog based lexical dictionaries to collect language resources [103]. Grammar rules on BEES translator are available to generate Sinhala words (Morphological generation rules) according to the given grammar.

Further, Silva et. al reported a prototype ESMTS through the rule-based approach [104]. This system, capable of translating day-to-day work Sinhala sentences, into English. The system has only achieved a success rate of 75% with a corpus of 150 sentences.

In the region, there are several rule-based machine translation systems available Especially on Indian family languages[105]. Among them, A domain-restricted, English-Hindi MT system has been developed using the dependency parsing techniques with replacing the transfer phase of the classical analysis [106].

Rule-based systems are still useful to develop MT systems for low resources languages and languages has more grammatical representations. Further, rule-based systems associated with the few limitations. RBMT system requires a big dictionary to handle more translations. Further, it is difficult to incorporate all the required rules; therefore, some linguistic information still needs to be set manually. Further, handling idiomatic expression and ambiguity in a large scale rule-based machine translation is also difficult [107]. Note that, the rule-based systems still challenge to adapt to a new domain.

2.4.5 Example-based Machine Translation

Example based approach is one of the early approaches in the MT which was proposed in 1984 through English-Japanese abstract translation [108]. Example based machine translation systems (EBMTS) uses bilingual parallel corpora with sample sentences on both languages to training. Thus the main activities of the EMBT can be categorized into three stages namely matching, retrieval and adaptation.

There are numbers of systems already developed through the example-based approaches including Japanese-English [109], Chinese-English [110] and Thai to Isarn [111]. Among others, Suhad and Yasir have been developed an EBMTS for English to Arabic [112]. This system uses two databases and introduces a solution to reduce redundancy problems.

EBMTS systems consist of advantages and limitations. These systems are more successful to translate phrasal verbs and easy to upgrade (insert examples to the database). However, EBMT systems are not successful when required examples are not available.

2.4.6 Statistical Approach to Machine Translation

The statistical approach is the most popular MT approach in the area of Machine translation [18]. The statistical approach generates translations using statistical methods through the bilingual text resources (Basically parallel corpus) [113]. Figure 2.9 shows the basic activity and task of the statistical MT system.

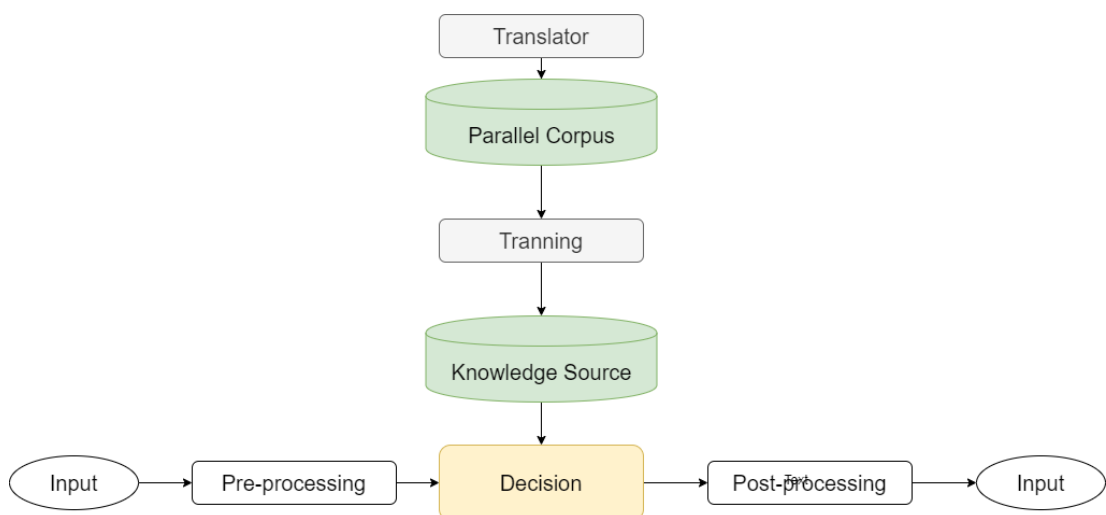


Figure 2.9: Activities on statistical MT

Note that, a statistical translation system requires language model probabilities, translation probabilities and a method for searching among possible source sentences. According to that, the statistical translation can be a model with equation 2.1. In most cases, the initial model of the statistical machine translation is based on “Bayes Theorem”, proposed by Brown [114] [115].

$$\Pr(S | T) = \frac{\Pr(S)\Pr(T|S)}{\Pr(T)} \quad (2.1)$$

where:

$\Pr(S|T)$ - Translation model

$\Pr(T)$ – Language mode

During the last few decades, a large number of Statistical MT systems have been developed: including Moss, Babel Fish, Bing and Google Translator.

Moses is one of the open-source, statistical MT systems that easily applies to the related language pairs [116]. The Moses system also consists of phrase-based and tree-based translation models to provide accurate translations. Besides, the system uses factored translation models to integrate linguistic and other information at the word-level [117]. The latest Moses consists of an Experiment Management System that allows for using Moses easily.

Babel Fish [118] is an online web page or text translator developed by AltaVista. Babel Fish uses Google translation service, which is provided by Google [48]. However, the Google translation service provides more service (Number of languages) than the Babel Fish [119].

Microsoft Translator is one of the commonly used multilingual MT tool provided by Microsoft since 1999 [120]. This translator uses four different approaches for MT to powering up the accuracy of the translation namely neural machine translation, Syntax-based and Phrase-based statistical machine translation. At present Microsoft translator support for more than 71 languages and integrated across multiple products.

Google Translator is free most popular MT system that supports more than 100 languages[121]. At the early stages, Google translator uses statistical approach and recently it moves to Neural Machine Translation [122]. This translator uses statistical or patterns analysis algorithms to deliver its translation. Hence, the system does not think of grammatically correct translation [123]. Still google translation comes with few limitations including words limits in the translation and does not always provide the grammatically correct translation

Few statistical machine translation systems were developed for the Sinhala language, especially Sinhala and Tamil. Among others, Pushpananda et. al have developed a statistical machine translation system for Sinhala-Tamil [124]. Besides, Rajpirathap and others have developed a Sinhala-Tamil statistical machine translation system using the Sri Lankan parliament corpus [125].

Note that, some of the current research investigates that the quality of the statistical machine translation can be improved using a rule-based phrase-level generation [126] especially including morphological analyser and generator for providing grammatically correct output [127].

Furthermore, Ranatunga and others have been developed a Statistical MT system named Si-Ta. This system should capable to translate Sinhala and Tamil government documents without loss of semantics [128].

Still, the statistical approach for MT is the most-used MT approach when neural machine translation is more popular. However, there are few challenges available in the statistical machine translation, including corpus creation and translation of non-related language pairs (with different word order).

2.4.7 Neural Machine Translation

Neural machine translation (NMT) can be considered as a successful approach to machine translation that uses machine learning concepts[19][129]. There are numbers

of language models already used for the neural machine translation, including the recurrent neural language model, feed-forward neural language model, long short-term memory models, deep models, and neural translation models. These neural translation model uses the encoder-decoder approach or adding an alignment model for more accurate translations. Besides, most of the researchers move to convolutional neural networks to develop their language model.

At present Google's Neural Machine Translation [122] system and TensorFlow's Neural Machine Translation model [130] are the most successful machine translation models that are capable of providing a more accurate solution for the phrase-by-phrase machine translation.

Thang Luong and others have developed a neural machine translation (NMT) model (TensorFlow's NMT model), which is based on a sequence-to-sequence (seq2seq) model [131]. This model can read the entire source sentence, understand its meaning, and then produce a translation through the deep recurrent architecture.

The NMT system first reads the source sentence using an encoder to build a thought vector, a sequence of numbers that represents the sentence meaning; a decoder then processes the sentence vector to emit a translation, as illustrated. NMT consists of two components: an encoder [132], which computes representations for each source sentence and a decoder, which generates one target word at a time and hence decomposes the conditional probability. Figure 2.10 shows the encoder-decoder architecture of the neural machine translation by TensorFlow. This NMT addresses the local translation problem in the traditional phrase-based approaches.

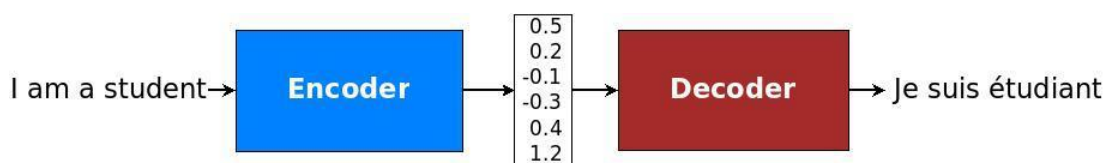


Figure 2.10: Encoder-decoder architecture of the NMT

Google's NMT is a popular Machine translation approach that uses end-to-end learning approach for automated translation. At present, Google's NMT supports more than 59 languages [133]. This NMT model consists of "a deep "LSTM network"(Long short-term memory network) with eight encoder and eight decoder layers using residual connections".

OpenNMT is an open-source toolkit for neural machine translation [134] that prioritizes efficiency, modularity, and extensibility to support NMT research. The OpenNMT toolkit consists of several facilities including modelling and translation support, summarization and, image-to-text, or speech-recognition. OpenNMT consists of several new features including the general-purpose interface, easily configurable models, and training procedures.

Stanford NLP group [118] has also developed a neural machine translation (NMT) model. This model has been tested with languages English-German and English-Czech [135] This model uses the attention-based neural machine translation method for better translation.

Pasidu and others have developed an NMT system for Sinhala and Tamil to translate among the official government documents [136] using Bahdanau's NMT architecture [137]. Through this research, they attempt to investigate NMT for small parallel corpus.

Neural machine translation systems still have some challenges [138][139]. Any machine translation system train for different domains is a critical task. However, the neural machine translation system can handle it easily (capable of domain adaptation). Further, same as statistical MT, neural MT systems require more data for training and testing. Thus, it takes some difficulties to build a successful neural machine translation system for low resources languages like Sinhala. (For instance, English-Spanish systems on WMT consist of 385.7 million English words paired with Spanish). Further, handling long sentences and word alignment is more difficult when both languages are non-related language pairs.

2.4.8 Knowledge-based Approach

Knowledge-based Machine Translation (KBMT) is an early approach for MT, that much popular in 1990-2000 years [16][140] [141]. In general, most of the KBMT system comprises of an ontology of concepts, analysis lexical and grammars, generation lexical and grammars and mapping rules between the Interlingua to provide an accurate translation.

KBMT-89 project at Carmegie Mellon University's Center for Machine Translation is the historical development for KBMT[142]. Further, Knowledge-based Accurate Natural-language Translation (KANT) is one of the successful, large-scale, commercial quality, domain-specific knowledge-based MT system for French, German and Japanese translations[143].

2.4.9 Transfer-based Machine Translation

In the transfer-based MT, systems take source language analysis results as a language depending model and transfer into a target language with considering word phrase or syntax [144]. Therefore, transfer-based systems can be classified into three groups namely word-based, phrase-based and syntax-based.

The Phrase-based model translates phrases as atomic units [26]. This model is the most used method in Machine translation before the neural Machine translation introduce. This phrase-based model can be used to translate the local context in a sentence easily. Still, some Google translators also use this model. In addition to that, some grammar checkers also uses this phrase-based model for more accurate automatic grammatical error correction [145].

This approach is used by the many machine translation systems with the support of the other approaches, including statistical [14], neural and rule-based approaches.

Google phrase-base machine translation is a more useful translator in the world, especially for low resource languages. Google translation starts with a rule-based system then move to phrase-base machine translation, now they already move to NMT.

Another phrase-based NMT system [146] has been developed using SleepWAKE Networks (SWAN) [147] that was recently proposed segmentation-based sequence modelling method. This system has been tested with English-German and English-Vietnamese machine translations.

Joshua is an open-source, statistical phrase-based MT system [148] toolkit of human languages. This system introduces a “phrase-based decoder” that uses the “standard priority-queue-based decoding algorithm” [149] to construct a hypergraph.

In the Indian region, few phrase-based machine translations [150] are available, PanchBhoota is a domain-specific phrase-based MT system for five Indian languages [151]. This system was specially developed for the three-domains, namely Health, Tourism and General domain.

A syntax-based MT system has been developed for English to the Hindi language [152]. This system uses Statistical and syntax-based approach to incorporate the representation of the syntax of the source into the statistical system. Note that, most of the transfer based MT systems use Statistical methods and the translation (or model) is based on the syntax, phrase or word-based [153] [154].

2.4.10 Agent-based Approach to Machine Translation

An agent-based approach for MT is a modern and powerful way to build MT systems by using the power of agents’ communication and parallel processing. In early days, few numbers of MAS systems available for NLP with different agent models including multi-agent systems based text understanding and clustering system for the car insurance domain [155] and TALISMAN [156]. However, these systems are not directly related to machine translation. A few years back, numbers of language processing systems are already developed through the MAS technology.

Aref and others have developed the Natural language understanding system using MAS technology through the lexical structural approach and a cognitive structural

approach [157]. The system takes user input and understands the input through the agents' communication.

Chunqi et al. have developed MAS with translation agents [158] to promote the efficiency in MT by shifting the paradigm from the transparent-channel metaphor to the human-interpreter metaphor. According to the agent architecture, analyze the interaction process and describes the decision.

Recently, Multi-agent technology has been used to improve the training process of the Conventional Neural Machine Translation [159]. Though the agent communication agent sharing the required knowledge. In this system, agents work together and increase the performance and quality of the translation.

However, still, there is no any MT system available for use of complete MAS concepts or approach for machine translation.

2.4.11 Hybrid Approach to Machine Translation

Machine translation can be done by considering two or more approaches. These systems can be categorized as a hybrid machine translation system. Nowadays, most of the machine translation system uses two more approaches to enhance translation quality than the usage of the single approach [20]. Most of the machine translation system based on rule-based (Hybridization guided by RBM) or corpus-based (Hybridization guided by corpus-based MT).

There is several Machine translation system available with hybrid concepts. Note that, most of the present statistical MT system powered with some other approaches (RBMT or NMT) [14].

Kiril and others have been developed Deep Machine Translation to English to Bulgarian with hybrid approach [160]. This system uses pre-processing and post-processing modules as well as two-level transfer. In addition, two two-level transfer methods, WordNets [161] for both languages are used as language resources. This MT

system also comprises a predominantly statistical component (factor-based SMT in Moses) [162] with some focused rule-based elements.

Hindi-English data-driven MT has been developed using SMT, EBMT and RBMT approaches [163]. The translation process of the system is based on rule-based MT.

In here, most of the Machine translation system has been discussed except English - Sinhala Machine translation. The reset section briefly shows some existing English Sinhala Machine Translation systems.

2.5 Local Resource and Existing ESMTS

Numbers of local language resources are developed since 1990 [164]. Among them, University of Colombo School of Computing (UCSC) [165] have developed various Sinhala language tools including Sinhala copra, Corpus-based Sinhala Lexicon [166], Sinhala WordNet [167], Sinhala Speech Recognition system [168] and Sinhala and Tamil Machine translation systems [169]. In addition to that, the University of Moratuwa has been conducted a number of Sinhala language research to enhance the Sinhala resources in the web, including SinMin - A Corpus for the Sinhala Language [170], Sin-Ta -Sinhala-Tamil Translator [171], Sinhala PoS tagged data set [172], EnSiTip [173], Sinhala WordNet [167], Morphological Analyzer [174] etc.

Considering the English Sinhala machine translation system availability, there are few numbers of research have been done. Historically, Weeresinghe's Sinhala to Tamil MT system using "corpus-based approach" [175] [176], and Vithanage's English to Sinhala rule-based MT systems for weather forecasting [177] can be considered as the historical developments. Also, Fernando and others developed ESMTS using Artificial Neural Networks [178]. Another research conducted to develop a prototype rule-based machine translation system for English to Sinhala and vice versa (SEES) [179]. This system takes Singlish as input and translates according to the existing rules. BEES is also English to Sinhala Machine Translation System, that follows a direct transfer approach [180]. This system uses computational grammar for Sinhala word

generation through the concept of Varanageema (Conjugation) [181]. The BEES has been implemented as a web-based and standalone application with Java and PROLOG for language processing [182].

2.6 Some Issues in Machine Translation

Machine Translation has been considered as a research challenging task so far. The below section describes some of these practical MT issues briefly.

2.6.1 Word and Sentence Segmentation

Word and sentence segmentation are essential before the translation process. Word and sentence segmentation are done through the computer system; text segmentation is the process of dividing the written text into words, sentences, or topics to take the correct meaning. However, some other languages like Thai, the boundary of the sentence is fuzzy [183] [184]. In the Sinhala language, limited numbers of research are conducted for voice [185] and text. The basic algorithm used to segment text at the end of a line or end word with dot sign (sentence termination symbol) [186].

2.6.2 Word Conjugation

Word conjugation (generate multiple word forms) [187] is another challenge in machine translation, especially when the target language is morphologically rich. To address these issues, the MT system needs an accurate “word generator” or “morphological generator” to generate appropriate word form for the given grammar. Considering the requirements of the EnSiMAS, “Sinhala morphological generator” can be used to handle word conjugation requirements [180].

2.6.3 Tense Detection

Tense detection [188] is another issue in MT. Tenses of the language and the sentence patterns are different from languages. For instance, in the English language, there are 12 tenses for active and 8 for passive voice [189]. However, the Sinhala language, there are only 3 tenses (present, past, and future) [190]. Therefore, appropriate tense needs to be identified through syntactical analysis, and it is required to generate appropriate target language tense through the target language syntax generation.

2.6.4 Multi-word Expression

Multi-word-Expression [191] (MWE) is an expression which made two or more words can be syntactically or semantically appear as a single unit) [192]. Machine Translation point of view it is required to identify those units together for the accurate translation; otherwise, translation results become meaningless.

2.6.5 Out of Vocabulary

Out-of-vocabulary is a term used to explain the input which is not present in a system's dictionary or database [193]. To solve these out-of-vocabulary issues number of methods are available including spelling expansion, morphological expansion, dictionary term expansion or proper name transliteration [194]. This can be applied for many machine translations, including English to Sinhala machine translation.

2.6.6 Translating Idiomatic Phrases

The idiomatic phrase gives an entirely different meaning. There is a no systematic way to take the given meaning of the idiomatic expression. Thus Idiomatic phrase translation has been still considered as a research challenge.

2.7 Summarization of Existing MT Approaches

Based on the above review and categorization, existing systems, advantages and limitations on each approach has been summarized in table 2.1

Table 2.1: Summary of the selected MT approaches

#	Approach	Example	Features	Limitation
1	Interlingua	ICENT [66], UNITRAN[69] English-Hindi interlingua-based machine translation [70], Sanskrit to English[73]	Easy to introduce new language	An analysis is more complex (All the levels of analysis is required)
2	Human- assisted	Anusaaraka (Among Indian Languages) [81] OmegaT [78]	humans and machines co- operate	Semi-automated
3	Dictionary- based	Pali to Sinhala MT [91] Czech and Russian [92]	Only required a bilingual dictionary	Success only for related languages
4	Rule-Based	BEES [99] Toshiba[97]	Provides grammatically correct translation	More rules than more complex
5	Example- based	English-Japanese [109] English-Arabic [112] Chinese-English [110]	Easily to update	Without examples translation is difficult
5	Statistical	Moses [116] Si-Ta [1128] Google Translator [121]	Most used approach with high accuracy	Sometimes result has some grammatical issues
6	Neural Machine Translation	Google Translator [121] OpenNMT [134] Stanford NLP [135] Sinhala-Tamil [136]	More accurate	More training and parallel corpus required

7	Knowledge-based	KBMT-89 [142] KANT [143]	A bit closer to human translation	Knowledgebase creation is difficult
8	Phrase-based	SWAN [147] Joshua [148]	More accurate than word-based	More rules than more complex
9	Agent-based	No complete MAS system for MT. TALISMAN [56]	Handle complexity in Natural languages	Difficult to implement and model
10	Hybrid	Moses [116] Hindi-English data-driven MT [163]	Provide better accuracy	Difficult to implement

2.8 Problem Definition

In the above, many machine translation approaches have been discussed. However, all of these machine translation systems should be unable to archive quality of the translation same as the human-generated translation. Thus, there is a quality gap between human translation and machine translation.

2.9 Summary

This chapter provides a critical review of MT, including existing approaches and systems. Compared with the existing MT approaches, the neural machine translation is the modern approach to machine translation, which provides some ability to reduce the quality gap between human translation and machine translation. However, it requires a sufficient amount of large parallel corpus. Note that, Sinhala language is a low resource language and required such resources are not available yet. Compared with human translation, all other existing machine translation approaches have some limitations and gives less translation quality. Hence there is a quality gap between human translation and machine translation.

CHAPTER 3

LITERATURE REVIEW AND BACKGROUND

3.1 Introduction

The previous chapter provided a review of machine translation, including approaches, systems, and issues. This chapter gives fundamental knowledge of English and Sinhala languages, including morphology syntax and semantics, which are required to build a machine translation system.

3.2 Computational Grammar for the English Language

The English language is the global communication west Germanic language, originate from the Anglo-Saxons family [195]. Then Modern English consists of 26 letters including five vowels [189]. This section briefly reports a comparatives study on Morphology, Syntax and semantics on the English language, with considering the requirements of the MT.

3.2.1 The Morphology of the English Language

The internal structure of a word can be described through the morphology. Theoretically, words are build-up from morphemes which is called as the smallest meaning bearing unit of a language [84]. For instance, the English word dog consists of a single morpheme, and the English word dogs comprise of two morphemes ‘boy’ and ‘s’. Note that, Compared with other morphologically rich languages like Sinhala, Sanskrit, few rules are available for the English. Table 3.1 shows the common suffixes available for English [196].

Table 3.1: Some Suffixes in English

Affix	Grammatical Category	Mark	Part of Speech
-s	Number	plural	nouns
's/'s	Case	genitive	nouns and noun phrases, pronouns
-self	Case	reflexive	pronoun
-ing	Aspect	progressive	verbs
-en/-ed	Aspect	perfect non- progressive	verbs
-ed	Tense	past (simple)	verbs
-s	Person, Number, Aspect, Tense	3rd person singular present	verbs
-er	Degree of Comparison	comparative	adjectives (monosyllabic or ending in -y or -i.e.)
-est	Degree of Comparison	superlative	adjectives

English Noun Morphology

The English noun is the main morphological category of the English language, which participates in inflexion and derivation [196] “Inflexion is the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood” [197]. English nouns participate in number, gender, and case inflexion. Table 3.2 shows some morphological rules for English noun inflexions.

Table 3.2 Some Morphological rules for Noun Inflection

Morphological Rules for English Nouns				
Grammar	Morphology			Example
	Base form	Add	Remove	
Singular	Noun	-	-	School
Plural	Noun	s	-	Schools
Plural	Noun	es	-	Dishes
Plural	Noun	ies	y	Ladies
Plural	Noun	ves	f	Wives
Singular Possessive	Noun	's	-	Book's
Plural Possessive	Noun	s'	-	Girls'
Singular	Verb	er	-	Reader
Plural	Verb	ers	-	Readers
Singular	Verb	ment	-	Achievement
Plural	Verb	ments	-	Achievements

According to the English noun inflexion, a noun can be divided into a regular noun or an irregular noun. Table 3.3 shows some regular and irregular noun forms.

Table 3.3 Regular and irregular Noun forms

Inflexion form	Regular	Irregular
Singular	computer	corpus
Plural	computers	corpora
Singular Possessive	computer's	corpus's
Plural Possessive	computers'	corpora's

English Verb Morphology

English verbs show few morphological forms when compared with the Sinhala language, namely the simple present tense, third-person singular, simple past tense, present participle, and past participle. According to the verb, conjugated English verbs can also be categorized into regular and irregular forms. Table 3.4 shows the regular and irregular verb forms.

Table 3.4: Regular and irregular English verb forms

Inflexion form	Regular	Irregular
Infinitive	help	write
Simple present	helps	writes
Present Participle	helping	writing
Past	helped	wrote
past Participle	helped	written

According to the English noun-verb inflexion morphology, a verb can be divided into a regular verb or irregular verb. Table 3.5 shows some rules for verb conjugation.

Table 3.5: Some Morphological rules for Verb conjugation

Morphological Rules for English Verb				
Grammar	Morphology			Example
	Base form	Add	Remove	
Infinitive	Verb	-	-	help
Simple present	Verb	s	-	helps
Present Participle	Verb	ing	-	helping
Past	Verb	ed	-	helped
past Participle	Verb	ed	-	helped

English Adjective Morphology

Adjectives describe the quantity, qualities, and/or states of a noun in other words. Adjectives can modify the noun. In addition to that, adjectives also appear as a complement to linking verbs or the verb. According to the degrees of comparison of an adjective, three forms are available, namely absolute, comparative, and superlative. Table 3.6 shows the morphological rules available for English adjectives.

Table 3.6: Adjective relationship of a Noun

Morphological Rules for English Adjective				
Grammar	Morphological			Example
	Base	Add	Remove	
(Positive) Adjective	Base	-	-	Bad
(Positive) Adjective	Noun Base	ish	-	Boyish
(Positive) Adjective	Noun Base	ful	-	Useful
(Positive) Adjective	Noun Base	less	-	Shameless
(Positive) Adjective	Noun Base	en	-	Golden
(Positive) Adjective	Noun Base	active	-	Talkative
(Positive) Adjective	Noun Base	able	-	Moveable
(comparative) Adjective	Adjective	er	-	Cleaner
(comparative) Adjective	Adjective	r	-	Larger
(comparative) Adjective	Adjective	ier	y	Dirtier
(Superlative) Adjective	Adjective	est	-	Cleverest
(Superlative) Adjective	Adjective	st	-	Simplest
(Superlative) Adjective	Adjective	iest	y	Dirtiest

English Adverb Morphology

An adverb is a kind of word modifier that modifies an adjective, or a verb in a sentence. In general, adverbs are ended with the suffix -ly [198]. In addition, Adverbs gives a full description of “how something happens”, using When, How, Where, “in what way” and “to what extent”. Table 3.7 shows some relation between verb and adverb.

Table 3.7: Verb and adverb usage

Verb	Adverb	Example
When?	early	She always arrives early.
How?	carefully	He drives carefully.
Where?	everywhere	They go everywhere together.
In what way?	slowly	She eats slowly.
To what extent?	slowly	It is slowly hot

3.2.2 Syntax of the English Language

Words can be arranged in different orders to create meaningful sentences. This arrangement has a grammatical structure called syntax. Compared with the Sinhala language, English has its grammatical structure. According to English grammar, an English sentence can be categorised into four main groups, namely declarative, interrogative, imperative, and conditional [199]. In general, each sentence has two major components, such as a subject and predicate.

The Subject

The subject is one of the main parts of the “sentence that performs an action. According to the structure, “the subject can be divided into simple or compound. The simple subject may be a noun phrase or a nominative personal pronoun”. The following context-free grammar shows the selected grammar rules that demonstrate the English subject.

⟨Subject⟩ → ⟨Simple Subject⟩

⟨Subject⟩ → ⟨Compound Subject⟩

⟨Simple Subject⟩ → ⟨Noun phrase⟩

⟨Simple Subject⟩ → ⟨Nominative Personal Pronoun⟩

⟨Compound Subject⟩ → ⟨Simple Subject⟩⟨ Conjunction⟩ ⟨Simple Subject⟩

⟨Noun Phrase⟩ → ⟨Article⟩ ⟨ specific proper noun⟩

⟨Noun Phrase⟩ → ⟨ Proper Noun⟩

⟨Noun Phrase⟩ → ⟨ Non-personal Pronoun⟩

⟨Noun Phrase⟩ → ⟨Article⟩ ⟨ Noun ⟩

⟨Noun Phrase⟩ → ⟨Article⟩⟨ Adjective ⟩ ⟨ Noun ⟩

⟨Noun Phrase⟩ → ⟨Article⟩⟨ Adverb ⟩ ⟨ Adjective ⟩ ⟨ Noun ⟩

⟨Noun Phrase⟩ → ⟨ Noun ⟩

⟨Noun Phrase⟩ → ⟨ Adjective ⟩ ⟨ Noun ⟩

⟨Noun Phrase⟩ → ⟨ Adverb ⟩ ⟨ Adjective ⟩ ⟨ Noun ⟩

The English Predicate

The predicate is also another part of the sentence that modifies the subject. Thus the predicate must contain a verb with or without a modifier. Therefore, the predicate can be divided into a verb or verb phrase and a subject. The following context-free grammar shows the selected grammar rules that demonstrate the English predicate.

⟨Predicate⟩ → ⟨Verb⟩ ⟨Complement⟩

⟨Predicate⟩ → ⟨Verb Phrase⟩ ⟨Complement⟩

⟨Verb⟩ → ⟨Simple present ⟩

⟨Verb⟩ → ⟨ Infinitive ⟩

⟨Verb ⟩ → ⟨Linking Verb⟩

⟨Verb Phrase ⟩ → ⟨Aux Verb⟩⟨ Infinitive ⟩

⟨Verb Phrase ⟩ → ⟨Aux Verb⟩ ⟨ NOT ⟩⟨ Infinitive ⟩

The following context-free grammar shows the selected grammar rules that demonstrate the complement

⟨Complement ⟩ → ⟨Ving ⟩

⟨ Complement ⟩ → ⟨to⟩⟨ Vinf ⟩ ⟨ Simple Object⟩

⟨ Complement ⟩ → ⟨Preposion phrase(s)⟩

⟨ Complement ⟩ → ⟨Adverb⟩⟨ Adjective ⟩

⟨ Simple Object ⟩ → ⟨Noun Phrase⟩

Verb Tense

Compared with a noun, a verb shows more forms of inflexion. A tense is an inflexion form of a verb that is used to describe the time variation of the action. The following context-free grammar shows the selected grammar rules that demonstrate the tense.

⟨Simple Present⟩ → ⟨ V3ps ⟩

⟨Simple past ⟩ → ⟨VPast⟩

⟨Simple Future ⟩ → ⟨Will ⟩⟨Vinf ⟩

⟨Present Continuous⟩ → ⟨ am, is, are ⟩⟨ V+ing ⟩

⟨Past Continuous⟩ → ⟨ was, were ⟩⟨ V+ing ⟩

⟨Future Continuous⟩ → ⟨ will ⟩⟨ be ⟩⟨ V+ing ⟩

3.2.7 The Semantics of English Language

Machine translation system point of view, it is essential to analyse all the levels of language representations. Semantic processing can be considered as the upper middle-level language processing method that represents meaning. Meaning can be defined by word-level, phrase-level, sentence-level, and paragraph-level [200].

Word-level Semantics

A particular word has one or more meanings. This is called word-level semantics. For instance, the word “book” has two meanings, namely a collection of a document (as a noun) or reserved some (as a verb). This word-level ambiguity can be solved by considering a phrase or sentence (other levels of semantics).

Phrase-level semantics

One or more words join together with some structure (particular orders) to make the phrase. Same as words, phrases generate separate meaning that is called phrase-level semantics. For instance, the phrase “will book” has a different meaning than its individuals.

Sentence Level Semantics

Sentence-level semantics refers to the meaning that deals with the meaning of syntactic units[201]. The thematic relationship is used to build the meaning of a sentence considering the relationship between each noun phrase and the verb phrases. Table 3.8 shows the thematic relation in a sentence.

Table 3.8: Basic Thematic Relationship in a sentence

Relationship	Description
Agent	Performs the action
Experiencer	Receives sensory or emotional input
Recipient	A special kind of goal associated with verbs expressing a change in ownership, possession
Theme	Undergoes the action but does not change its state
Patient	Undergoes the action and changes its state
Location	Where the action occurs
Source or Origin	Where the action originated
Time	The time at which the action occurs
Beneficiary	The entity for whose benefit the action occurs

Pragmatics (The paragraphs/text Level Semantics)

Rather than a single sentence, pragmatics takes the “context contributes to meaning” machine translation point of view. Pragmatics analysis is a solution for the “word sense

ambiguity”[202]. To identify paragraph-level semantic, it requires taking complete knowledge of each existing sentences in the selected paragraph.

3.3 The Sinhala Language

The word Sinhala (සිංහල) is derived from the lion (සිංහ) and taker (ල). The Sinhala language is one of the official languages of Sri Lanka. Sinhala has its own alphabet with few versions, including the Unicode version (18 vowels and 45 consonants) and alphabet of the “Sedath sagara” (10 vowels and 20 consonants) [190]. However, Sinhala mix alphabet consists of 18 vowels and 36 consonants. Note that Sinhala consists of only four parts-of-speech (පද) Namely Noun (නාම), Verb (ක්‍රියා) and Nipath (නිපාත) and Updarga (indeclinable particle). This four parts-of-speech embrace the eight parts-of-speech specified in English. The following figure shows the part of speech mapping between English and Sinhala.

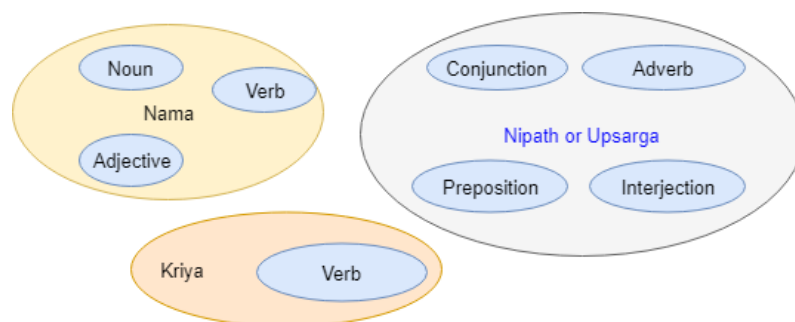


Figure 3.1: Part of speech mapping between English and Sinhala

3.3.1 Morphology of the Sinhala Language

Sinhala is an inflationary rich language, and it participates in inflexion, derivation, and conjugation of nouns and verbs. The next section briefly describes a noun and verb morphology of the Sinhala language.

Sinhala Noun Morphology

Sinhala Noun (*Nama*) represents a Noun, ProNoun, and Adjective in the English language. Sinhala is also an inflectionally rich language and shows a strong relationship between subject and verbs. Also, Sinhala Noun shows Gender, Number, Person and Case-based inflexion. For Instance, more than 27 Noun formas can generate by using a Sinhala Noun. (9 cases for each Singular direct, Singular indirect and plural form). Table 3.9 shows Sinhala Noun inflexion forms for the Sinhala word මුවා (dear)

Table 3.9: Sinhala Noun inflexion form for base word මුවා (dear)

Case	Singular direct	Singular indirect	Plural
Nominative	මුවා	මුවෙක්	මුවෝ
Accusative	මුවා	මුවකු	මුවන්
Instrumental	මුවා විසින්	මුවකු විසින්	මුවන් විසින්
Auxiliary	මුවා ගෙන්	මුවකු ගෙන්	මුවන්ගෙන්
Dative	මුවාට	මුවකුට	මුවන්ට
Ablative	මුවාගෙන්	මුවකුගෙන්	මුවන්ගෙන්
Genitive	මුවාගේ	මුවකුගේ	මුවන්ගේ
Locative	මුවා කෙරෙහි	මුවකු කෙරෙහි	මුවන්කෙරේ
Vocative	මුවා		මුවනේ

Sinhala Verb Morphology

A Sinhala verb is the action word in a sentence that can be divided into two groups, namely transitive and intransitive. In general, Sinhala verbs are inflected from five categories, namely voice, mood, tense, number, and person. Compared with the English verb, the Sinhala verb takes only three tenses such as the present, past, and future; however, the Sinhala verb shows more inflexion (verb conjugation) than the Sinhala noun. More than 36 inflexion forms, including active, passive, optative mood,

imperative mood, and conditional mood are available for a Sinhala base verb. Table 3.10 shows some inflexion forms for the Sinhala verb “මරණවා”. Table 3.11 shows the add-remove values for the Sinhala verb for the verb conjugation. Note that these rules can be applied for agent-based verb generation.

Sinhala verb is the action word in a sentence that can be divided into two groups, namely transitive and intransitive. In general Sinhala, verbs are inflected from 4 categories, namely voice, mood, tense, number and person. Compare with the English verb Sinhala verb takes only three tenses such as the present, past and future; however, Sinhala verb shows more inflexion (verb conjugation) than Sinhala Noun. More than 36 inflexion forms including active, passive, optative mood, imperative mood and conditional mood available for a Sinhala base verb. Table 3.10 shows some inflexion forms for the Sinhala verb “මරණවා”. Table 3.11 shows the add-remove values for the Sinhala verb for the Verb conjugation. Note that, these rules can be applied for agents’ based verb generation.

Table 3.10: Verb inflexion forms for Verb *Maranawa* (මර ධාතු ව)

කාලය	උත්තම ඒක	උත්තම බහු	මධ්‍යම ඒක	මධ්‍යම බහු	ප්‍රථම ඒක	ප්‍රථම බහු
කතෘ වර්තමාන	මරමි	මරමු	මරහි	මරහු	මරයි	මරති
කතෘ අනාගත	මරන්නෙමි	මරන්නෙමු	මරන්නෙහි	මරන්නෙහු	මරන්නේ	මරන්නෝ
කතෘ අතීත	මැරිමි	මැරූමු	මැරිහි	මැරූහු	මැරී	මැරූ
කර්ම වර්තමාන	මැරෙමි	මැරෙමු	මැරෙහි	මැරෙහු	මැරෙයි	මැරෙති
කර්ම අනාගත	මැරෙන්නෙමි	මැරෙන්නෙමු	මැරෙන්නෙහි	මැරෙන්නෙහු	මැරෙන්නේ	මැරෙන්නෝ
කර්ම අතීත	මැරිණිමි	මැරූණුමු	මැරිණිහි	මැරූණුහු	මැරිණි	මැරූණු

Table 3.11: Add-remove values for the Sinhala verb

Rule	Base Verb	Active voice Present Tense	Active voice past Tense	Active past tps	Active present tpp	Active Futue	Passive voice Present Tense	Passive voice past Tense	Passie past tps	Passive present tpp	Passive Futue
201	ගැනීම	ග	ගත්තෙ	ගත්තේය	ගත්තෝය	ග	ගනුලබ	ගනුලැබුවෙ	ගනුලැබුවේය	ගනුලැබුවෝය	ගනුලැබ
202	නීම	න	නුවෙ	නුවේය	නුවෝය	න	නනුලබ	නනුලැබුවෙ	නනුලැබුවේය	නනුලැබුවෝය	නනුලබ
203	නීම	නි	නුවෙ	නුවේය	නුවෝය	නි	නිනුලබ	නිනුලැබුවෙ	නිනුලැබුවේය	නිනුලැබුවෝය	නිනුලබ
204	නෑම	නෑ	නෑවෙ	නෑවේය	නෑවෝය	නෑ	නෑනුලබ	නෑනුලැබුවෙ	නෑනුලැබුවේය	නෑනුලැබුවෝය	නෑනුලබ
205	ණෑම	ණෑ	ණෑවෙ	ණෑවේය	ණෑවෝය	ණෑ	ණෑනුලබ	ණෑනුලැබුවෙ	ණෑනුලැබුවේය	ණෑනුලැබුවෝය	ණෑනුලබ
206	ණීම	ණි	ණිතෙ	ණිතේය	ණිතෝය	ණි	ණිනුලබ	ණිනුලැබුවෙ	ණිනුලැබුවේය	ණිනුලැබුවෝය	ණිනුලබ
207	රීම	ර	ලෛ	ලෛය	ලෛෝය	ර	රනුලබ	රනුලැබුවෙ	රනුලැබුවේය	රනුලැබුවෝය	රනුලබ
208	රීම	රි	රියෙ	රියේය	රියෝය	රි	රිනුලබ,	රිනුලැබුවෙ	රිනුලැබුවේය	රිනුලැබුවෝය	රිනුලබ
209	කෑම	ක	කෑවෙ	කෑවේය	කෑවෝය	ක	කනුලබ	කනුලැබුවෙ	කනුලැබුවේය	කනුලැබුවෝය	කනුලබ
210	කීම	කි	කූතෙ	කූතේය	කූතෝය	කි	කිනුලබ	කිනුලැබුවෙ	කිනුලැබුවේය	කිනුලැබුවෝය	කිනුලබ
211	ඟෑම	ඟෑ	ඟෑවෙ	ඟෑවේය	ඟෑවෝය	ඟෑනුලබ	ඟෑනුලබ	ඟෑනුලැබුවෙ	ඟෑනුලැබුවේය	ඟෑනුලැබුවෝය	ඟෑනුලබ
212	ගේම	ගේ	ගත්තෙ	ගත්තේය	ගත්තෝය	ගේ	ගේනුලබ	ගේනුලැබුවෙ	ගේනුලැබුවේය	ගේනුලැබුවෝය	ගේනුලබ
213	ටීම	ට	ටුවෙ	ටුවේය	ටුවෝය	ට	ටනුලබ	ටනුලැබුවෙ	ටනුලැබුවේය	ටනුලැබුවෝය	ටනුලබ
214	ටීම	ටි	ටුවෙ	ටුවේය	ටුවෝය	ටි	ටිනුලබ	ටිනුලැබුවෙ	ටිනුලැබුවේය	ටිනුලැබුවෝය	ටිනුලබ
215	ඩීම	ඩ	ඩුවෙ	ඩුවේය	ඩුවෝය	ඩ	ඩනුලබ	ඩනුලැබුවෙ	ඩනුලැබුවේය	ඩනුලැබුවෝය	ඩනුලබ
216	ඩීම	ඩි	ඩියෙ	ඩියේය	ඩියෝය	ඩි	ඩිනුලබ	ඩිනුලැබුවෙ	ඩිනුලැබුවේය	ඩිනුලැබුවෝය	ඩිනුලබ
217	තෑනීම	ත	තෑනුවෙ	තෑනුවේය	තෑනුවෝය	ත	තනුලබ	තනුලැබුවෙ	තනුලැබුවේය	තනුලැබුවෝය	තනුලබ
218	කීම	කි	කූවෙ	කූවේය	කූවෝය	කි	කනුලබ	කනුලැබුවෙ	කනුලැබුවේය	කනුලැබුවෝය	කනුලබ
219	දීම	ද	දූවෙ	දූවේය	දූවෝය	ද	දනුලබ	දනුලැබුවෙ	දනුලැබුවේය	දනුලැබුවෝය	දනුලබ
220	දීම	දි	දූවෙ	දූවේය	දූවෝය	දි	දිනුලබ	දිනුලැබුවෙ	දිනුලැබුවේය	දිනුලැබුවෝය	දිනුලබ
221	බීම	බ	බූවෙ	බූවේය	බූවෝය	බ	බනුලබ	බනුලැබුවෙ	බනුලැබුවේය	බනුලැබුවෝය	බනුලබ
222	යෑම	ය	ගියෙ	ගියේය	ගියෝය	ය	යනුලබ	යනුලැබුවෙ	යනුලැබුවේය	යනුලැබුවෝය	යනුලබ
223	පෑම	පෑ	පෑවෙ	පෑවේය	පෑවෝය	පෑ	පෑනුලබ	පෑනුලැබුවෙ	පෑනුලැබුවේය	පෑනුලැබුවෝය	පෑනුලබ
224	හීම	හ	හෙ	හත්තේය	හත්තෝය	හ	හනුලබ	හනුලැබුවෙ	හනුලැබුවේය	හනුලැබුවෝය	හනුලබ
225	මීම	ම	මුවෙ	මුවේය	මුවෝය	ම	මනුලබ	මනුලැබුවෙ	මනුලැබුවේය	මනුලැබුවෝය	මනුලබ
226	වීම	ව	වූයෙ	වූයේය	වූවෝය	ව	වනුලබ	වනුලැබුවෙ	වනුලැබුවේය	වනුලැබුවෝය	වනුලබ
227	හීම	හි	හින්	හින්තේය	හින්තෝය	හි	හිනුලබ	හිනුලැබුවෙ	හිනුලැබුවේය	හිනුලැබුවෝය	හිනුලබ
228	රීම	රි	රියෙ	රියේය	රියෝය	රි	රිනුලබ	රිනුලැබුවෙ	රිනුලැබුවේය	රිනුලැබුවෝය	රිනුලබ
229	දීම	දෙ	දුන්නෙ	දුන්නේය	දුන්නෝය	දෙ	දෙනුලබ	දෙනුලැබුවෙ	දෙනුලැබුවේය	දෙනුලැබුවෝය	දෙනුලබ
230	සීම	ස	සූවෙ	සූවේය	සූවෝය	ස	සනුලබ	සනුලැබුවෙ	සනුලැබුවේය	සනුලැබුවෝය	සනුලබ
231	ටීම	ටි	ටින්තෙ	ටින්තේය	ටින්තෝය	ටි	ටිනුලබ	ටිනුලැබුවෙ	ටිනුලැබුවේය	ටිනුලැබුවෝය	ටිනුලබ
232	දීම	දෙ	දෙන්නෙ	දෙන්නේය	දෙන්නෝය	දෙ	දෙනුලබ	දෙනුලැබුවෙ	දෙනුලැබුවේය	දෙනුලැබුවෝය	දෙනුලබ
233	දීම	ද	දන්නෙ	දන්නේය	දන්නෝය	ද	දනුලබ	දනුලැබුවෙ	දනුලැබුවේය	දනුලැබුවෝය	දනුලබ
234	ලීම	ල	ලන්නෙ	ලන්නේය	ලන්නෝය	ල	ලනුලබ	ලනුලැබුවෙ	ලනුලැබුවේය	ලනුලැබුවෝය	ලනුලබ
235	වීම	වෙ	වෙන්නෙ	වෙන්නේය	වෙන්නෝය	වෙ	වෙනුලබ	වෙනුලැබුවෙ	වෙනුලැබුවේය	වෙනුලැබුවෝය	වෙනුලබ
236	ඵීම	ඵ	ඵන්නෙ	ඵන්නේය	ඵන්නෝය	ඵ	ඵනුලබ	ඵනුලැබුවෙ	ඵනුලැබුවේය	ඵනුලැබුවෝය	ඵනුලබ

3.3.2 Syntax of the Sinhala Language

Same as the English language, Sinhala has its syntax, and it differs from English. According to the structure, a Sinhala sentence can be classified into six subcategories, namely simple, complex, constructed, collectral, compound, and illectical. The simple sentence consists of a Sinhala subject and a single finite verb. According to Sinhala grammar, a Sinhala sentence can be divided into eight components, namely attributive adjunct of the subject, a subject attributive adjunct of an object, object, attributive adjunct of the predicate, attributive adjunct of the complement of a predicate, the complement of a predicate, and predicate. In addition to that, more than 40 syntax rules are available to make a correct Sinhala sentence, including subject-verb agreement. The following context-free grammar shows the selected grammar rules that demonstrate the Sinhala sentence structure.

$$\langle \text{Sentence} \rangle \rightarrow \langle \text{Subject Phrase} \rangle \langle \text{Verb Phrase} \rangle$$
$$\langle \text{Sentence} \rangle \rightarrow \langle \text{Subject Phrase} \rangle \langle \text{Object Phrase} \rangle \langle \text{Verb Phrase} \rangle$$
$$\langle \text{Object Phrase} \rangle \rightarrow \langle \text{Object} \rangle$$
$$\langle \text{Subject Phrase} \rangle \rightarrow \langle \text{Subject} \rangle$$
$$\langle \text{Verb Phrase} \rangle \rightarrow \langle \text{Verb} \rangle$$
$$\langle \text{Subject Phrase} \rangle \rightarrow \langle \text{Attributive adjunct of Subject} \rangle \langle \text{Subject} \rangle$$
$$\langle \text{Object Phrase} \rangle \rightarrow \langle \text{Attributive adjunct of Object} \rangle \langle \text{Object} \rangle$$
$$\langle \text{Subject} \rangle \rightarrow \langle \text{Noun} \rangle$$
$$\langle \text{Object} \rangle \rightarrow \langle \text{Noun} \rangle$$

3.4 Comparison Between English and Sinhala Languages

Sinhala and English languages take some similarities and many differences. Table 3.12 presents a summary in between English and Sinhala languages.

Table 3.12: Fundamental differences in both Sinhala and English

Category	English	Sinhala
Alphabet	There are 5 vowels and 21 consonants	There are 18 vowels and 42 consonants in this modern alphabet [203]
Part of Speech	There are eight parts of speech in the English language:	There are 4 Nama, Kriya, Nipatha and Upsarga
literary Morphology	minimal	Noun (case, number, definiteness and animacy) Sinhalese mark person, number or gender on the verb
Verbal morphology	there is few subject-verb agreement	there is more subject-verb agreement (PNG)
Syntax	SVO (subject-verb-object) word order Left-branching language	SOV (subject-object-verb) word order Left-branching language
Discourse		Sinhalese is a pro-drop language

3.5 Summary

This chapter described more details on English and Sinhala languages with more attention to morphology and syntax in both languages. The next chapter presents natural language processing techniques that are required to build a machine translation system.

CHAPTER 4

NATURAL LANGUAGE PROCESSING TECHNIQUES

4.1 Introduction

The previous chapter discussed in more details on English and Sinhala languages with more attention on morphology, syntax and semantics on both languages. This chapter gives detail about some existing natural language processing techniques, including Language modelling, morphological analysis, syntax analysis, morphological and syntactical generation.

4.2 Computational Model for English and Sinhala

According to the existing grammar in both languages, computational model required to handle both language information for the translation. Therefore, a relational model has been designed to represent knowledge in both languages. According to the developed model basic unit of the language is a word. Thus the model has been based on the English and Sinhala word. Note that, an English phrase has several English words, a sentence consists of several phrases and a paragraph consist of several sentences. With the above basic concepts, the object-oriented model has been designed with composite objects. This model consists of four knowledge representation levels, namely words with its morphological knowledge, phrases with syntax knowledge, a sentence with thematic relationships and paragraph with pragmatics knowledge. Figure 4.1 shows the knowledge representation level of the English language.

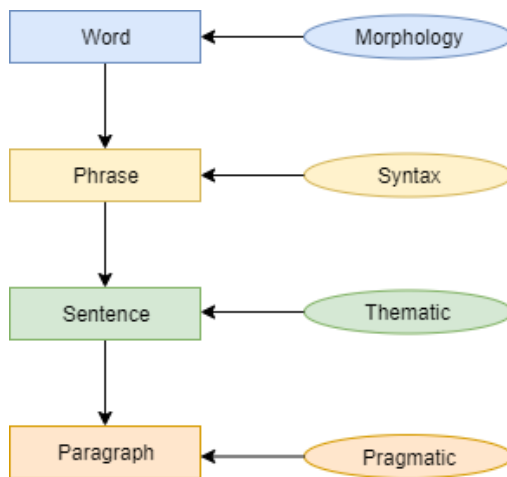


Figure 4.1: Language model for English and Sinhala

With the above idea, an ontological model has been designed and developed for the English and Sinhala languages. This ontological model is based on word, phrase and sentence.

Further, as we know the English language consists of 8 part of speech, including noun, verb, adjective, adverb and, etc. A word-based language model has been designed for English and Sinhala languages to represent language knowledge for the translation. Word level information for a word and some related semantics information also stored in the ontology. To represent a word, grammar, morphology and semantic features are stored in a word ontology including its identification tag. Figure 4.2 shows the ontology of a word.

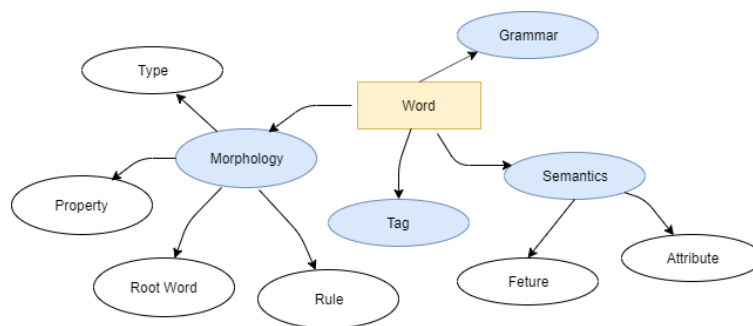


Figure 4.2: Ontology of a word

Furthermore, the Sinhala language consists of 4 part of speech, namely Nama (noun), Kriya (verb), Nipatha (like preposition) and Upasarga. Same has English language

Ontology, word-based Sinhala language model has been designed for the Sinhala language to represent language knowledge for the translation. In addition to the above, English to Sinhala machine translation system requires to generate appropriate Sinhala word forms from the existing base word according to the given grammar. Thus, morphological generation rules are also stored in the Sinhala word ontology.

The phrase is made with one or more words. The ontological point of view each phrase consists of a number of words, phrase type, headword information, location (Order of the phrase in a sentence) and thematic relationships. Figure 4.3 shows the ontological design of the English and Sinhala phrase.

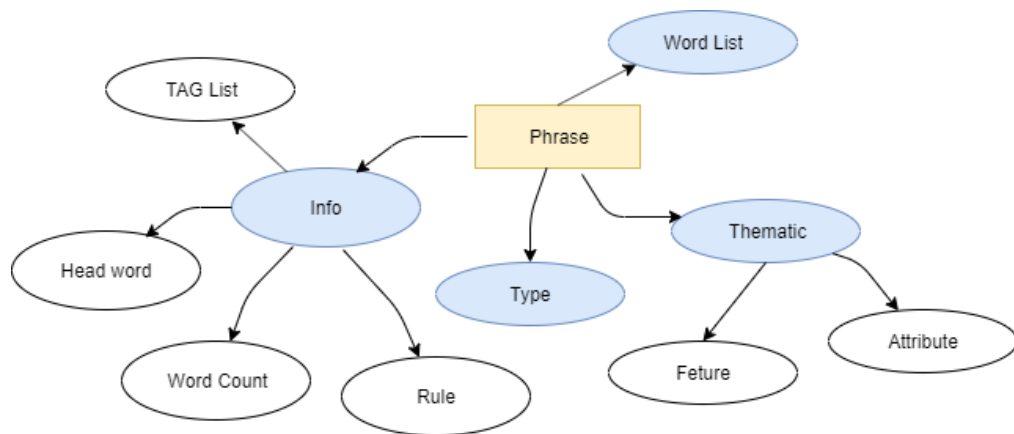


Figure 4.3: Ontology of a Phrase

A sentence is made with one or more phrases. The ontological point of view each sentence consist of number of phrases, sentence type, thematic information for the sentence. Figure 4.4 shows the ontological design of the English and Sinhala sentence.

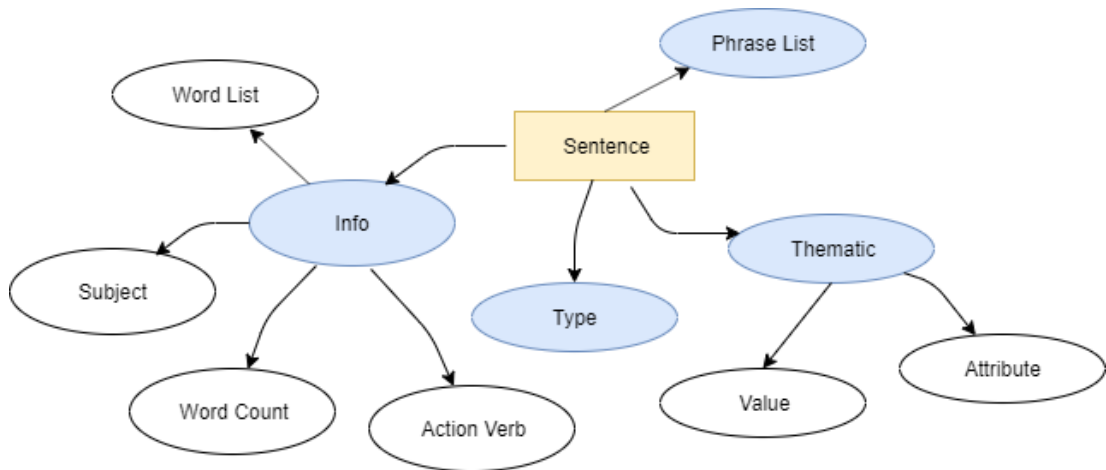


Figure 4.4: Ontology for a Sentence

With combining all the language modules, Object-Oriented model has been built to store sentence information which is useful to English to Sinhala machine translation. The next section briefly describes Morphological analysis and Morphological analysis and generation for machine translation.

4.3 Morphological Analysis and Generation

“Word” is the building block of the language. Machine translation point of view, it essential to identify words correctly before proceeding the process of machine translation. Morphological analysis is a word-level language analysis that identifies the internal structure of the word including the type of the word and the grammar. At present, there are numbers of approaches available for morphological analysis, including Morpheme-based, Lexeme-based and Word-based morphology [204]. Most of these Morphological analyzers have been developed as a part of the Machine translation system. There are very few researches have been conducted for the Sinhala Language morphological processing. Under the BEES project, a Sinhala Morphological analyzer has been developed [205]. The Analyzer uses Sinhala language morphological rules and a Sinhala word dictionary to take the analysis. Table 4.1 shows a summary of the existing metrological analyzers with their approach.

Table 4.1: Summary of the Existing Morphological analyzers

Language	Approach	Description
Tamil language	Morpheme-based	Lushanthan and others have developed [206] using Xerox toolkit and finite-state operations [207]
Hindi	Paradigm approach	part of the Hindi to Punjabi Language translation [208],
Marathi	Paradigm-based	Finite State Morphological analyzer [209] [210].
Pali	rule-based approach	David and others have been developed [211].
Kannada	paradigm approach	as a part of the English to Kannada Machine translation system [212].
Six Indian languages	Paradigm approach	Under the Anusaaraka system [213].
Hindi	uses finite-state transducers	Generic Morphological Analysis Shell [214].
English	Rule-based	Developed under the bees project [205]

Morphological generation is the backward approach of the morphological analysis that generates a correct word form for the given root word. According to the process, Morphological generation is a bit easy than the Morphological analysis. In general Morphological generator takes a base word and relevant grammar to generate the appropriate word form. Therefore, most of the systems capable of handling analysis and generation. Figure 4.5 shows the general pipeline of the Morphological analyzer and generator.

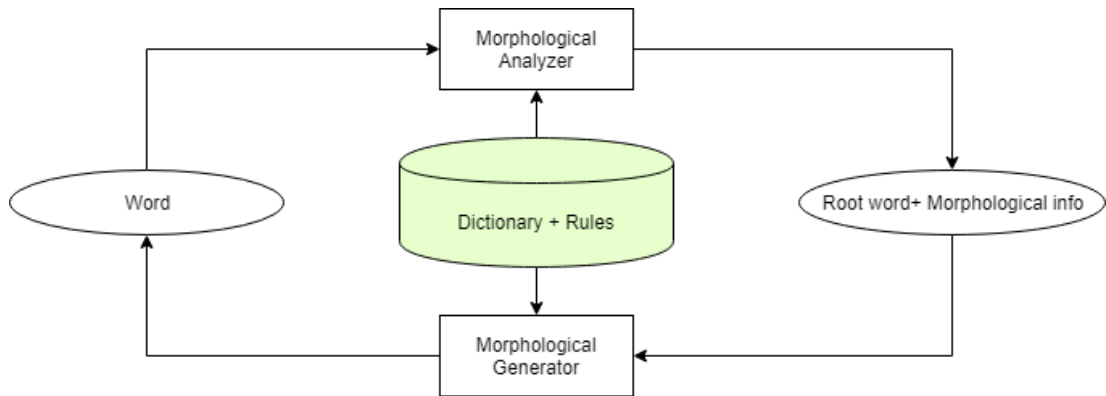


Figure 4.5: Process of the morphological analysis and generation

After the morphological analysis, the system should be able to take all the word information that can be used to process another level of language analysis, such as syntax and semantics.

4.4 Syntactical Analysis and Generation

A phrase is a meaning bearing unit in a sentence than a word. In other word, Phrase consists of numbers of words and each word in a phrase provide a collective meaning than the single word. Further, the structure of the sentence (syntax) can be analyzed through the Parsers. Parsers are computer application, capable to parse the sentence or a phrase through the given specific grammar. In general, parsers are classified through its parsing direction namely, top-down, bottom-up etc. For instance, top-down parsers are taken input from left-to-right and analyze the structure top to bottom. Several Top-down parsers are available including Recursive descent Parser [215] and CYK Parsers [216]. Further, there are numbers of tools available for Parser development including NLTK [217], OpenNLP [218], and JavaCC [219].

In addition to the above number of freely available successful parsers available for different languages, namely Stanford parser [220], The Link Grammar Parser [221], and Enju parser [222]. In addition to the above, a Prolog-based English parser has been already developed under the BEES project [223] [224].

Syntax generation is the opposite direction of the syntax analysis. Most of the syntax generation systems have been developed through rule-based methods. Syntax generator is required to generate grammatically correct target language syntax according to the given syntax.

Further, Phrase chunking is another type of application that segment a sentence into its sub constituents. In addition to the parsers, Noun phrase and verb phrase chunking is a commonly used method to identify noun phrase and verb phrases [225].

4.5 Semantics Processing

Semantics processing is a way to understand the meaning of the given context (sentence or phrase). Machine translation system point of view, it should be required to completely analyse the input and should able to generate or transfer the meaning into the target language correctly. In general, semantics processing can be divided into four levels, namely, word level, phrase level, sentence level and paragraph level. This section briefly describes semantics options for each case.

4.5.1 Word level semantics

Each word in the text consists of its meaning which is called word-level semantics. However this can be overwritten (may be replaced through the upper-level semantics) For instance, Take the phrase “The computer Society” if we take word individually computer refers to the computer and society has its meaning, however, as a phrase the word computer is in the adjective form and its change the direct meaning and only modify the property of the society. If we take some tense on the English sentence “will play”, “going to play” there is no direct meaning for the word “will” and “going”. According to the above facts, Machine translation point of view, it is required to consider word-level meaning as well as phrase level or sentence level meaning to take accurate translation.

4.5.2 Phrase level semantics

The phrase consists of one or more words each word consists of its meaning. However, if we consider a phrase, which has the main word which is used to deliver the meaning. Note that, in the previous example “Society” is the main word of the noun phrase, and other words are modifiers.

4.5.3 Sentence level semantics

The sentence consists of one or more phrases. Therefore, to take the sentence level semantics, it is required to extract the thematic relation on the sentence. The thematic relations give various roles for the sentence. There are numbers of thematic relations available, including Agent, Experiencer, Theme, Patient, Instrument, etc. In the machine translation point of view, thematic relation extraction and generation are beneficial to provide an accurate translation.

4.6 Summary

This chapter described an in-depth study on some related Natural Language Processing techniques for English to Sinhala Machine Translation including morphological processing, syntax processing and semantic processing. The next chapter discusses multi-agent technology and its features to model English to Sinhala machine translation system.

CHAPTER 5

MULTI AGENT TECHNOLOGY

5.1 Introduction

In the previous chapter, natural language processing techniques on machine translation were discussed, including morphological analysis, morphological generation, syntax analysis, syntax generation and semantics processing techniques. This chapter presents a critical study on multi-agent system technology and its features, including a review on the existing framework for multi-agent system development. Finally, this chapter also reports on MaSMT, a specially created framework for agent-based machine translation.

5.2 What is Multi-agent System?

A multi-agent system (MAS) is a computerised system composed of multiple interacting intelligent agents [226]. MAS consists of two or more agents, capable of communicating with each other in the shared environment. In general, a multi-agent system consists of four components: agents, environment, ontology and the virtual world. An agent may be a computer application or an independently running process (a thread) capable of doing some actions. Theoretically, agents are capable of acting independently (autonomous) and controlling their internal status according to the requirements. Agents communicate with other agents through messages. These messages consist of information for agent activities (agent doing their task according to the messages they have required from others). The environment is, to a large degree, the interaction between the “outside worlds” of the agent. In most of the cases, environments are implemented within a computer.

The Ontology [227] can be defined as an explicit specification of conceptualization. Ontologies capture the structure of the domain. Thus agent nature of the multi-agent system capabilities is based on the ontology. Further, considering the correct type of agent and making the agent communication are the most critical and essential

requirements of the agent-based system development. Further, multi-agent systems consist of several advantages [228]. Some of the common advantages are:

- A multi-agent system consists of interconnected agents; they used to distribute computational resources than the central resources. (However, most of the multi-agent systems run on a single machine.)
- A multi-agent system provides interconnection and interoperation of multiple systems.
- A multi-agent system should be capable of taking global coordinates, distributed information from sources efficiently.
- Multi-agent system solutions (optimal or accepted) for the current situation.

5.2.1 Type of Agents

According to the activities, behaviour, and existing features, agents can be categorised into different types including simple reflex agent, model-based reflex agent, goal-based agent, utility-based agent and learning agents. However, various classifications are also available for agents including, collaborative agent, interface agent, mobile agent, information or internet agent, hybrid agent and smart agent. These agents show different capabilities and behaviours to work together. Figure 5.1 shows a different view of agents.

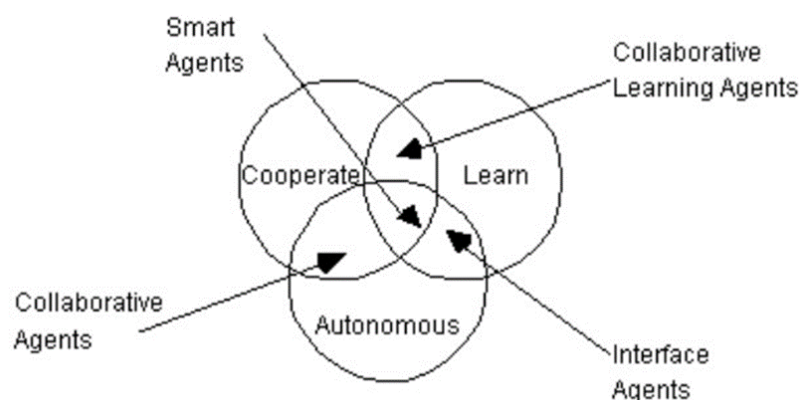


Figure 5.1: Different types of agents

5.2.2 Agent Communication

Communication among agents on MAS is the hidden factor for the success that allows collaboration among agents, negotiation and cooperation between agents, etc. Thus, agent communication needs a commonly understood semantics. Further, agents can communicate with each other with peer-to-peer, broadcast, or noticeboard methods.

In the peer-to-peer agent communication [229], an agent sends messages only for known peer(s). It means that there is an agreement between sender and receiver. The broadcasting method sends a message(s) for all. According to this method, everyone in the group, sometimes all agents in the system) require the message. The notice board method allows a different way than the above. The noticeboard method would send messages into the shared location (noticeboard). If the agent required some then agent can take that information from the noticeboard.

However, multi-agent systems take some time to make the solutions and that are required to pass messages among agents. The parallelism of the multi-agent system and avoiding unnecessary message parsing can help to support the fastest message parsing [230]. The multi-agent system is also used to handle the “system complexity” to provide intelligent solutions through agent communication [231]. Thus, multi-agent system development is also a bit of a complicated process when the system needs more communication. According to such complexity, selecting a suitable framework is highly important [232] than ad-hoc development. In general, a multi-agent framework provides agent infra-structure, communication, and monitoring methods for agents. In addition to that, common standards are available for agent development special agent communication, including FIPA-ACL [233] and KQML [234] [235]. FIPA is one of the common standards for agent development. A number of agent systems have been developed with the FIPA standard [236], including JADE.

5.3 Existing MAS Development Framework

Few multi-agent system development frameworks are available with different features. Among others, JADE [237] is a Java-based free and open-source software framework

for MAS development. JADE provides middle-ware software support with GUI tools for debugging and deployment. Further, JADE provides task execution and composition model for agent modelling, and peer-to-peer agent communication has been done with asynchronous message passing. In addition to the above, JADE consists of the following key features:

- JADE provides a FIPA-compliant distributed agent platform
- Multiple directory facilitator (change agents active at run time)
- Messages are transferred encoded as Java objects.

MaDKit is a Java-based multi-agent development platform, which is based on the AALAADIN conceptual model [238]. This organizational model consists of groups and roles for agents to manage different agent activities. MaDKit also provides a lightweight Java library for MAS design [239]. The architecture of MaDKit is based on three design principles, such as micro-kernel architecture, agentification of services, and the graphic component model. Also, MaDKit provides asynchronous message passing. Further, it can be used for designing any multi-agent applications, from distributed applications to multi-agent simulations[240].

PADE [241] is a free, entirely Python-based multi-agent development framework to develop, execute, and manage multi-agent systems in distributed computing environments. PADE uses the libraries from twisted project to allow communication among the network nodes. This framework support for multi-platforms, including embedded hardware that runs on Linux [242]. Also, PADE consists of some essential functionalities. PADE agents and its behaviours have been built using object-orientation concepts. PADE is capable of handling messages in FIPA-ACL standard and supports cyclic and timed behaviours.

SPADE (a Smart Python multi-Agent Development Environment) is a commonly used Python-based framework for multi-agent system development [243][244]. SPADE includes several features including the following: the SPADE agent platform is based

on the XMPP, support agent model based on behaviours, supports FIPA metadata using XMPP Data Forms, and provides a web-based interface for agent control.

Jason [245] is a fully Java-based, open-source, MAS development framework that provides speech-act based inter-agent communication. Jason has been developed using AgentSpeak [246]. The Jason framework consists of several features, including strong negation and annotations.

AgentBuilder [247] is a free quick agent and agent-based software development framework that is compatible with Java, as well as KQML and CORBA support. Several multi-agent systems have already been developed using AgentBuilder, including a “buying and selling system”.

“Shell for Simulated Agent Systems” (SeSAm) [248] is another MAS development environment that provides agent modelling facilities, as well as agent simulations [249]. SeSAm especially gives facilities to construct complex models. The SeSAm framework is used in many domains, including “logistics and production”.

The Agent Development Kit (ADK) [250] is a Java-based, open-source, MAS toolkit that allows quick-build, secure, large-scale solutions. ADK also uses agent-oriented G-net model including “end-to-end” tracing. Further, Table 5.1 shows a summary of the existing multi-agent system development frameworks.

Table 5.1: Summary of the existing Multi-agent system development frameworks

System	Type	Platform	Features
JADE	Open Source	Java	Asynchronous Message Parsing
MaDKit	Open Source	Java	Asynchronous Message Parsing,
PADE	Free	Python	Supports FIPA, cyclic and timed behaviours support
SPADE	Free	Python	Supports FIPA metadata using XMPP Data
AgentBuilder	Open-source	Java	Capable of building intelligent agent-based applications
Json	Open-source	Java	speech-act based inter-agent communication
SeSAm	Open-source	Programming Shell	GUI based agent modelling
ADK	Open-source	Mobile-based	Large-scale distributed solutions

5.4 MaSMT: Multi-agent Framework for Machine Translation

Existing multi-agent development frameworks directly does not support the distinct requirements of the proposed agent-based machine translation system. The proposed system requires several activities on natural language processing, including morphological processing, syntax processing, and semantic processing. Therefore, a new framework has been designed and developed, incorporating the following required features:

- Should able to handle more than 100 agents easily
- Provide the fastest message passing

- Agents can easily customise according to the requirements
- Agents should support local languages (Sinhala)

5.4.1 MaSMT Framework

The “Multi-Agent System for Machine Translation” (MaSMT) is a freely available Java-based MAS development framework, specially designed to implement EnSiMaS. The framework was first released in March 2016 as an open-source product. The rest of the section gives more details on inside the MaSMT.

5.4.2 AGR Organisational Model and MaSMT Architecture

The AGR organisational model was designed initially under the Aalaadin model, which consists of agents, groups, and roles. Figure 5.2 shows the UML-based Aalaadin model for multi-agent system development [251]. According to the model, each agent is a member of one or more groups, and a group contains one or more roles. The agent should be capable of handling those roles according to the agent’s requirements. This model is used by the MaDKit system by allowing free overlapping agents among groups [252].

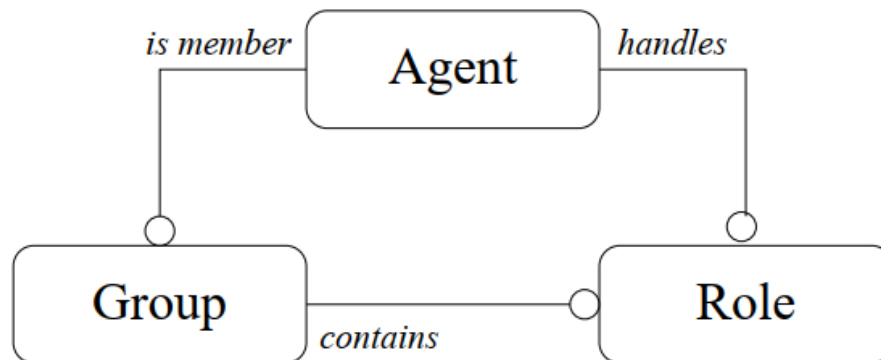


Figure 5.2: UML-Based Aalaadin model for multi-agent system development

Source: UML-based Aalaadin model [251].

The MaSMT model is almost the same as the above model but removes a freely overlapping feature of the group and role at the same time. It means the agent is a one member of one or more groups, as well as one or more roles; however; there is only one active group and role. Thus, the agent does actions according to this active group and role. Note that, agents are only active communicating entities capable of playing roles within groups. Therefore, MaDKiT provides the freedom for agent designers to design appropriate internal models for agents. With this idea, the MaSMT agent is designed considering the three-level architecture that consists of a root agent, controlling agents, and ordinary agents [253]. The MaSMT root agent contains several controlling agents (managers). Each controlling agent consists of any number of ordinary agents. In addition to that, agents can be clustered according to their group and role. This layered model allows for building agent swarms quickly. Fig. 5.3 shows the agent diagram of the three-level architecture on MaSMT.

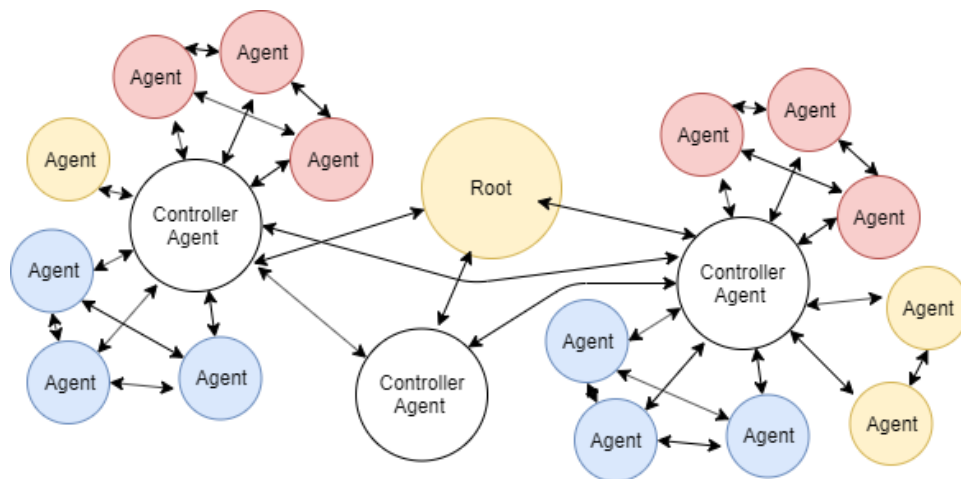


Figure 5.3: Agents' architecture on MaSMT

With this model, an ordinary agent can communicate with its swarm, as well as its controller agent. The controller agent should be capable of communicating and fully control its ordinary agents. Controllers can communicate with other controllers, and the root can handle all agents in the system. With this model, MaSMT allows for passing messages through the peer-to-peer, broadcast, or noticeboard methods.

5.4.3 MaSMTAbstractAgent

AbstractAgent model is used to identify agents through its group-rule-id. Agent identifier for the particular agent can generate using group, role and relevant id. Also, role(dot)id@group can be used to locate agents quickly. As an example read_words.101@ensimas.com provides read_words is a role, id is 101 and ensimas.com is a group.

This abstract agent model is used by the MaSMT to handle all the agent-based activities that are available in the MaSMT agents, MaSMT controllers, and MaSMT root agent.

5.3.4 MaSMT Agent

MaSMT agents are the active agents in the framework provides agents' infrastructure for agent development. The Modular architecture of the MaSMT Agent is shown in figure 5.4. MaSMT agent consists of several built-in features including Notice board reader and Environment controller.

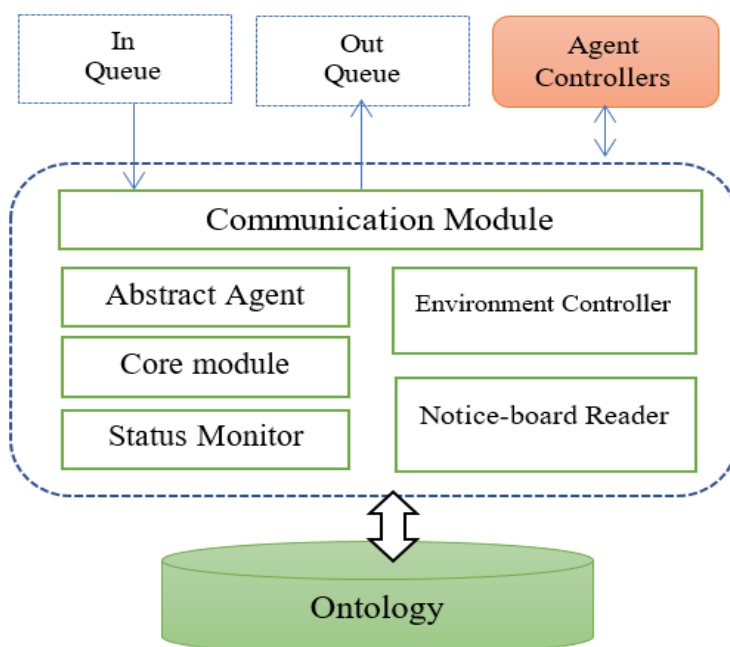


Figure 5.4: Modular architecture of the MaSMT Agent

5.3.5 MaSMT Agent's Life cycle

The MaSMT provides life cycle for each agent to handle its activities. Figure 5.5 shows the life cycle of the EnSiMaS agent.

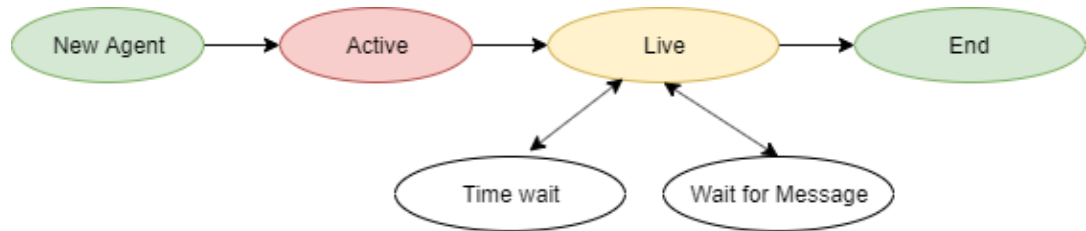


Figure 5.5: The life cycle of the MaSMT Agent

The MaSMT agent is a kind of Java-threaded program that consists of three sections, namely active, live, and end. More details are given in the MaSMT development guide.

5.3.6 MaSMT Controller agent

MaSMT controller agent is the middle-level controller agent of the MaSMT framework, capable to control its, clients, as required. Figure 5.6 shows the Architecture of the MaSMT controller agent. The MaSMT controller agent also provides all the features available in the MaSMT agents. In addition to that, MaSMT controller should capable to provide message passing, network access, notice board access and environment handling capabilities.

5.3.7 MaSMT Root Agent

The root agent is the top-level controller agent (MaSMT Manager), which is capable of handling other MaSMT controller agents. The MaSMT root agent is capable of communicating with other root agents through the “Net access agent”.

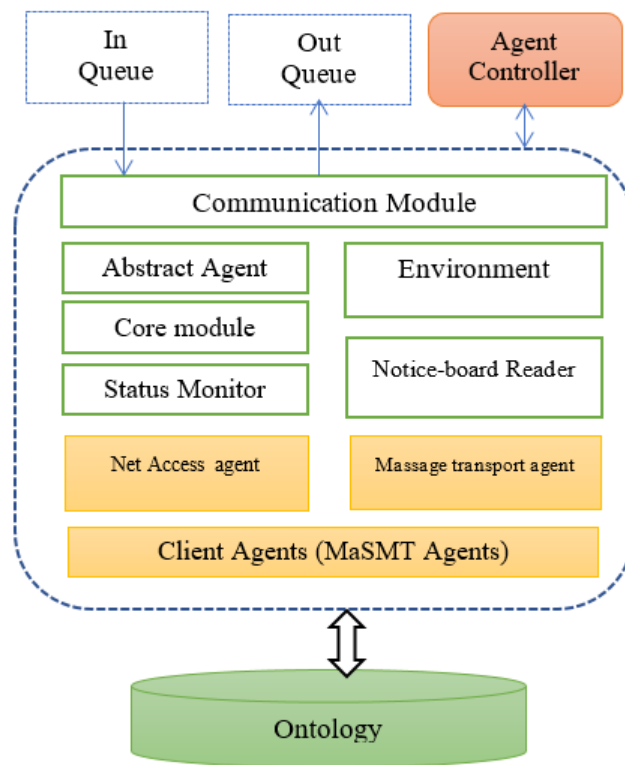


Figure 5.6: Architecture of the MaSMT controller agent

5.3.8 MaSMT Agents' Swarm

MaSMT allows two different levels of swarms, namely agent swarm and controller swarm, to build agent communication quickly. The root agent can handle the controller agent swarm, and the controller agent can handle MaSMT agent swarm. At present, any swarm can handle up to 2000 clients. Figure 5.7 shows the design of the MaSMT agent swarm. More details are present under the MaSMT guide.

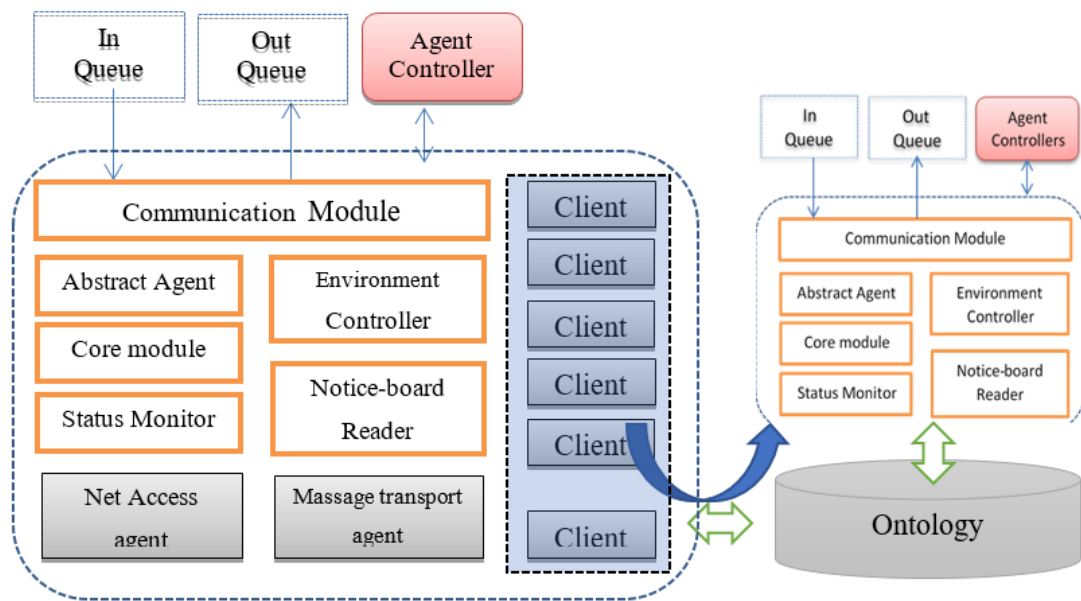


Figure 5.7: Design of the swarm of Agents

5.3.9 MaSMT Messages

The MaSMT framework uses messages named MaSMTMessages to provide agent communication. These MaSMTMessages have been designed using FIPA-ACL message standards. MaSMTMessages can be used to communicate in between MaSMT agents as well as other agents that support ‘FIPA-ACL’ message standards. Table 5.2 gives the structure of the ‘‘MaSMTMessage’’ including data fields and types. More information on ‘‘MaSMTMessages’’ is provided under the MaSMT development guide (Appendix B).

Table 5.2: Structure of the MaSMT Messages

Data Field	Type	Category
sender	MaSMTAbstractAgent	The sender of the message
receiver	MaSMTAbstractAgent	The receiver of the message
replyTo	MaSMTAbstractAgent	Who responsible for the conversation
message	String	Subject of the message
content	String	Content of the message
ontology	String	Ontological information
type	String	Type of the message
data	String	Some data (image, URL etc.)
header	String	direction of the message (MaSMT send message with considering it header)
language	String	Language and encoding type
conversationID	String	Conversation id of the message

5.3.10 MaSMT Settings

The MaSMT framework provides full customisation facilities to run MAS applications smoothly. For that, various settings need to change. Table 5.3 gives default values for the MaSMT.

Table 5.3: Default Settings of the MaSMT

Property	Description	Default value
Name	Name of the framework	MaSMT
version	The present version of the framework	3.0
Delay	Delay for each life circle (ms)	50
DebugMode	True or false (if true show report all agent's actions)	false
Proxy	Use to communicate with remote systems	0.0.0.0
Port	Use to communicate with remote systems	8088
AppPath	Path of the main system	""
Host	Machine name	localhost
MaXDelay	The maximum delay for wait for messages	100
MinDelay	Minimum delay for each agent's life circle	10

5.3.11 MaSMT Message Parsing

The MaSMT framework supports “peer-to-peer”, “broadcast”, “noticeboard approach” or email-based communication for message parsing. To send messages, it is required to follow few steps such as enable message parsing (use `activeMessageParsing` method), create the message and insert into the system through the “`addMessage`” or call “`broadCastToClients`” to send messages for all.

Read Messages

The MaSMT framework provides three methods to read messages, namely, `getMessage`, `peekMessage`, and “`waitForMessage`”. The “`getMessage`” method takes a message from its in-queue. The `peekMessage` method gets the front message from in-queue without changing its in-queue. The “`waitformessage`” method is more useful in the message parsing that waits until a message is on the “`inQueue`”.

Message Header

MaSMT messages can send according to its headers. Table 5.4 shows relevant headers that are used to send messages.

Table 5.4: Message directives (headers for messages)

Message Header	Message description
agents	Sends message(s) to the agent(s) who has given group and role
AgentGroup	Sends message(s) to an agent(s) who has given group
AgentRole	Sends message(s) to an agent(s) who has given a role
Local	Sends message(s) to it swarm who has given role and group
LocalRole	Sends message(s) to it swarm who has given a role

RoleOrGroup	Sends message(s) to agents with given “role or group.”
Broadcast	Sends message(s) to all members in its swarm
Root	Sends message(s) to Root agent
Controller	Sends message(s) to its Controller
NoticeBoard	Sends a message to the noticeboard
MailMessage	Sends a message as an email message

5.3.12 Applications of MaSMT

Using the MaSMT framework, various kinds of multi-agent systems have already been developed, including, Octopus [254], AgriCom [255] and RiceMart [256]. Octopus provides a multi-agent-based solution for Sinhala chatting; AgriCom and RiceMart provide a communication platform for the agricultural domain. Also, a multi-agent-based file-sharing application has already been developed for the distributed environment [257]. MaSMT is already developed as a Java application. Thus web-based application development capabilities have been already tested through the web-based event planning system [258].

5.4 Summary

This section briefly described an introduction to multi-agent system technology, including existing multi-agent frameworks. This section especially described the developed multi-agent development framework MaSMT. The agent model of the MaSMT, message parsing methods, and some existing development were also discussed in this chapter. The next chapter presents the proposed hybrid approach for MT.

CHAPTER 6

A HYBRID APPROACH TO MACHINE TRANSLATION

6.1 Introduction

This thesis proposes a psycholinguistic-based hybrid approach. The proposed approach was powered through psycholinguistic language translation concepts and multi-agent technology. In here, the psycholinguistic language translation method has been investigated by considering two-sentence processing models such as “the garden path model” [259] and “the constraint satisfaction model” [260]. In addition to that, it also considers four factors that are affected by human language parsing namely “the structure of the phrase”, “the thematic relationship of the phrase”, “the probability” and “the semantics features”. Further, the hypothesis noted in the thesis can be stated as “multi-agent system technology can be used to implement English to Sinhala Translation.” Therefore, a multi-agent system technology has been proposed to implement the proposed approach.

6.2 A Novel Approach to Machine Translation

The approach is based on the concept: “How can humans translate a sentence in a better way than the machine?” Therefore, psycholinguistic language translation techniques have been considered to model this proposed approach for machine translation. The proposed translation model has been implemented through the multi-agent system. Thus, this machine translation approach is stated as a “multi-agent approach for machine translation”. The next section presents the theoretical basis for the proposed approach.

6.3 Theoretical Basis of Language Translation

There are numbers of theories available for human-based language translation. Among others, the garden path model and the constraint satisfaction model are the psycholinguistic theories for language parsing. These two models demonstrate the concept of how people parse an English sentence together with the meaning.

The garden path model is one of the psycholinguistic models for language parsing that is restricted to a single context. The reader takes some context (meaning) for the first noun phrase and selects other phrases according to the previously selected meaning. Somehow, new information presented later, the existing context is not relevant for the new information. Then the reader should able to find different acceptable solutions for the existing context according to new information. For instance, the sentence “the horse raced past the barn fell” is a popular example that shows the said behaviour.

The constraint satisfaction model also states that the reader uses all the available information (including syntax, semantics, etc.) at once when engaging in the parsing of a sentence. Thus this model processes all the information in parallel. Note that, the proposed machine translation model (approach) has been designed considering the above two concepts. The approach uses a combination of the above-said models.

Also, human sentence parsing is based on phrase structure, probability, thematic roles, and semantic features. These four factors have some contribution to the translation. However, extract weight for the above four factors for the translation is still unknown. Figure 6.1 shows the four factors contributing to language parsing.

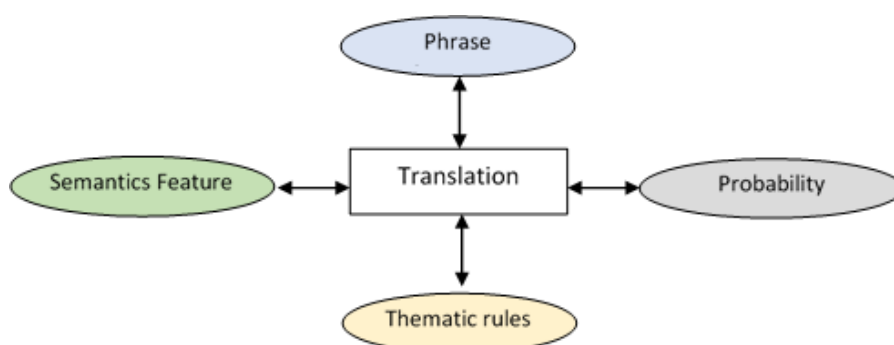


Figure 6.1: Factors contribute to sentence parsing

Phrase Structure

The structure of the phrase gives a grammatical relation in between each word in a phrase. There are different types of phrases available, including noun phrases, verb phrases, preposition phrases, etc. Considering the translation requirements, English phrases need to be converted to Sinhala. According to the structure of the two languages, these conversions need some rules. Note that, some English verb phrases consist of several English words but in Sinhala consist of different number. The English phrase ‘will eat’ translates into *Kmneya* (කන්නේය). Therefore, phrase transfer rules have been identified. Also, each phrase consists of a meaning bearing word (the headword) and its relationship with other words has been identified.

Probability

Each English word consists of several Sinhala translations. Thus, an English phrase has some Sinhala solutions. The probability can be used to select a suitable translation among existing solutions. In here, the probability of each Sinhala translation is calculated through Google Search Index (GSI). After creating multiple Sinhala phrases, usage of each phrase is calculated through the Google search index. The probability of a Sinhala phrase $P(x)$ is calculated through Equation 6.1.

$$\text{Probability Pro}(x) = \frac{\text{Usage of } P(x)}{\sum \text{Usage of } X} \quad (6.1)$$

where:

X is the usage of the Phrase P

Pro(X) is the probability of the Phrase P

Thematic roles

The thematic roles make the relationship in between each phrase in the sentence. The English syntax processing system is capable of taking these thematic relations [20].

Semantic feature

The semantic feature gives a clear meaning about words. However, the present system does not store the semantic features of a word. To avoid this limitation, EnSiMaS can provide multiple solutions.

6.4 A multi-agent Approach to Machine Translation

EnSiMaS consists of five language processing tasks, including EMA, ESA, SMG, SSG, and English-to-Sinhala phrase generation. Thus, multi-agent-based subsystems (swarms) have been developed to handle each required language processing activity.

6.4.1 Multi-agent Approach to English Morphological Analysis

Morphological analysis is a process to analyse the structure of words that are available in the input sentence. In here, English morphological agents are capable of analysing English morphology using a “rule-based root fixing method” to identify English words. For instance, an agent removes letter s if it is the last letter of the input word then searches the availability of the regular noun list. If it is available, then the original word is recognised as a plural noun. Same as if the new word is available in the regular verb list, then the original word is recognised as a simple present tense verb. Finally, the results are stored in the virtual world for further use.

6.4.2 Multi-agent Approach to English Syntax Analysis

“Syntax analysis” is the process of analysing the structure of the sentence or a part of a sentence. An agent-based English syntax analyser has been developed through the phrase-based chunking. Considering the existing phrase patterns, the available phrase has been recorded. Then the system should be capable of identifying some structure of the input sentence. Note that, the hypotheses proposed in the thesis are a human translation method to translate an English sentence into Sinhala. We argue that people

can identify some accepted meaning of an English sentence without much grammatical knowledge of English [261]. This system is capable of identifying existing phrases, with left-to-right sentence reading. These agents work as phrase chunking that separate and segment a sentence into its sub constituents, such as nouns, verbs, and prepositional phrases. Then the active phrase modification agent reads all the phrases and identifies valid phrases on the available phrase list. Then the "active phrase searching agent" and the "sentence agent" are capable of analysing a syntactic structure for the input. Then "thematic agent" is capable of identifying existing thematic rules for each phrase in the input sentence.

6.4.3 Multi-agent Approach to English to Sinhala Phrase-based Translation

The proposed approach is based on how humans translate the given English phrase into Sinhala. According to the approach, each English phrase translates into several Sinhala phrases considering the said four factors effected to the language parsing. According to the sentence structure, suitable Sinhala translated phrases are selected considering the thermostatic relations and the probability for each case. The approach is powered through rule-based and agent-based machine translation techniques.

6.4.4 Multi-agent Approach to Sinhala Morphological Generation

Compared with English, Sinhala is a morphologically rich language. Thus, EnSiMaS required an accurate Sinhala word generation module to provide grammatically correct translations. In general, morphological generation (generates appropriate word form for the given base word and the grammar) required related morphological rules to generate relevant word forms. Therefore, existing morphological rules are taken from the BEES project [180]

6.4.5 Multi-agent Approach to Sinhala Syntax Generation

There are numbers of syntactical differences available in the English and Sinhala languages. Therefore, Syntax generation is an essential task for the EnSiMaS, that provides a grammatically correct translation. This Sinhala syntax generator uses a set of syntax transfer rules to re-order the existing phrase list.

6.5 Why a Multi-agent Approach?

Compared with existing technologies for machine translation, it is of high demand to develop a human-quality MT system for English to Sinhala. The current trend to develop machine translation systems is to add one or more intelligent techniques to improve the quality of machine translation. However, the human translation process is unknown, and that gives the best translation solution so far. Thus, the theoretical basis of the approach is used as psycholinguistic parsing techniques that are used in human language processing. To model these theories, the multi-agent approach is used.

The multi-agent approach is one of the modern approaches that are capable of handling complexity through the agents' communication. In addition to the above, multi-agent systems are capable of providing intelligence solutions than the existing approaches. Also, multi-agent models can quickly be adapted to simulate human behaviours because they are an intelligent software development technique.

With the above ideas, this thesis presents a multi-agent approach to implement human translation behaviour through the agents. The implemented system is named as EnSiMaS, which is capable of translating English sentences into Sinhala through a hybrid approach. This approach originates from psycholinguistic phrasing, which is based on humans' language translation.

6.6 Features of EnSiMaS

EnSiMaS consists of the following key features.

- EnSiMaS provides phrase-based and agent-based strategies for English to Sinhala sentence-based translations.
- EnSiMaS has been designed with a theoretical basis of psycholinguistic language parsing.
- The lexical resource of EnSiMaS is stored in the lexical databases, and object-oriented models are also used to store language ontology.
- EnSiMaS can be used as a standalone application.
- EnSiMaS has been implemented with Java. Therefore, it is platform-independent.
- EnSiMaS comprises of three language translation tools, namely EnSiMaS Dictionary, EnSiMaS Phrase-based Editor, and EnSiMaS Translator.
- EnSiMaS provides multiple translations.

6.7 Input for EnSiMaS

As mentioned in the above, EnSiMaS comprises of three different language translation tools. According to that, it accepts different types of inputs. The EnSiMaS Dictionary takes an English word as an input. EnSiMaS Phrase-based Editor takes grammatically correct English sentences, and the translator takes any text (number of sentences).

6.8 Output of EnSiMaS

EnSiMaS provides multiple solutions, considering the probability of usage. Thus, EnSiMaS Dictionary provides suitable Sinhala words for the given English word according to the usage of the Sinhala term. The EnSiMaS Phrase-based Editor provides

grammatically correct Sinhala phrases for intermediate editing. EnSiMaS Translator provides multiple Sinhala translations.

6.9 Process of the EnSiMaS

The proposed hybrid approach has been simulated with the following steps.

1- Creates a new Ontology

As the first step of the translation, a new ontology is created with all the available facts.

2 – English Morphological analysis

It is required to analyse English morphology for each English word. Therefore, it takes the morphological analysis and results are stored in the virtual world.

3 – Semantics Mapping

Takes available Sinhala terms for the existing English words from the bilingual dictionary. Results are also stored in the virtual world.

4 - English syntax analysis

Identifies the syntax of the input sentence. In here, all the available English phrases and the thematic relationship on each phrase have been taken. Also, thematic relations for the input sentence are recorded.

5- Creates phrase agents

Generates relevant Sinhala phrase agents for the existing English phrases

6 – Phrase Translation

Multiple Sinhala translations are taken for the existing English phrase using “phrase structure” and the “thematic relationship” of each phrase.

7 – Select suitable Sinhala phrases using calculated probability

The Sinhala phrase agent calculates the probability values for each phrase and identifies the most suitable Sinhala solution. Using “subject-verb”, “object-verb”, and “subject-object-verb” agreement, suitable Sinhala phrases are selected according to their calculated probability.

8 – Sinhala Syntax Generation

The Sinhala syntax processing system reads the existing syntax of the input sentence and re-arranges the Sinhala phrases to generate the grammatically correct Sinhala sentence. Further, the thematic relationship is also recorded for further use.

6.10 Summary

This chapter described a proposed hybrid approach for machine translation considering “the garden path model”, “the constraint satisfaction model”, and psycholinguistic parsing concepts as the theoretical bases of the approach. This chapter also discussed the translation process, input, output, and features of the approach.

CHAPTER 7

DESIGN OF THE ENSIMAS

7.1 Introduction

The previous chapter proposed a hybrid approach, including the theoretical basis of the MT approach, hypothesis of the research, input, output, and process of the translator. This chapter reports the design of the EnSiMaS.

7.2 Design of the EnSiMaS

The proposed hybrid approach has been simulated through the EnSiMaS. Thus EnSiMaS capable of translate English text (sentence(s)) into Sinhala. Figure 7.1 shows the design diagram of the EnSiMaS.

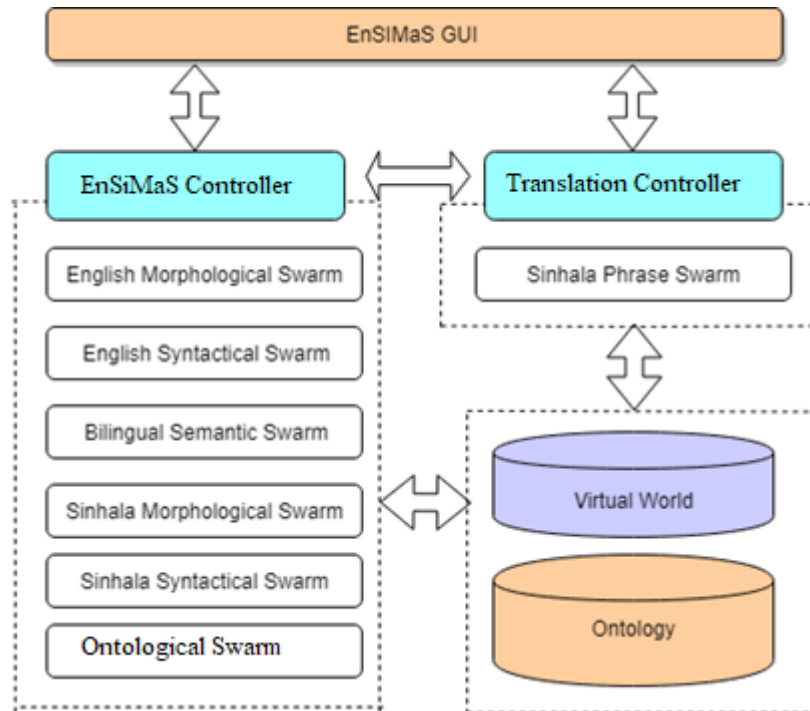


Figure 7.1: Design of the EnSiMaS

EnSiMaS has been modelled with 12 modules including GUI, EnSiMaS controller, Translation controller, five language processing supporting modules, ontological swarm, virtual world, and ontology. The EnSiMaS controller handles all the language processing activities on the machine translation. The translation controller handles the Sinhala phrase swarm to manage phrase agents.

7.2.1 EnSiMaS GUI

The EnSiMaS GUI is the user interface of the system that takes English text (sentence(s)) and provides translated Sinhala text (Sentence(s)). EnSiMaS GUI is the main application container of the system should capable to load and handle all other agents including MaSMT controller and translation controller.

7.2.2 Ontology

To achieve meaningful translation, the translation system requires all levels of language information (Morphological to Pragmatics) on both source and target. For that, English lexical information, English-Sinhala bilingual information, and Sinhala lexical information are models to store inside the ontology. In addition to the above, morphological and syntactical rules are also modelled to store inside the ontology. Therefore, this ontology consists of English lexical information, English-Sinhala bilingual information and Sinhala lexical information. Figure 7.2 shows the design of the EnSiMaS Ontology.

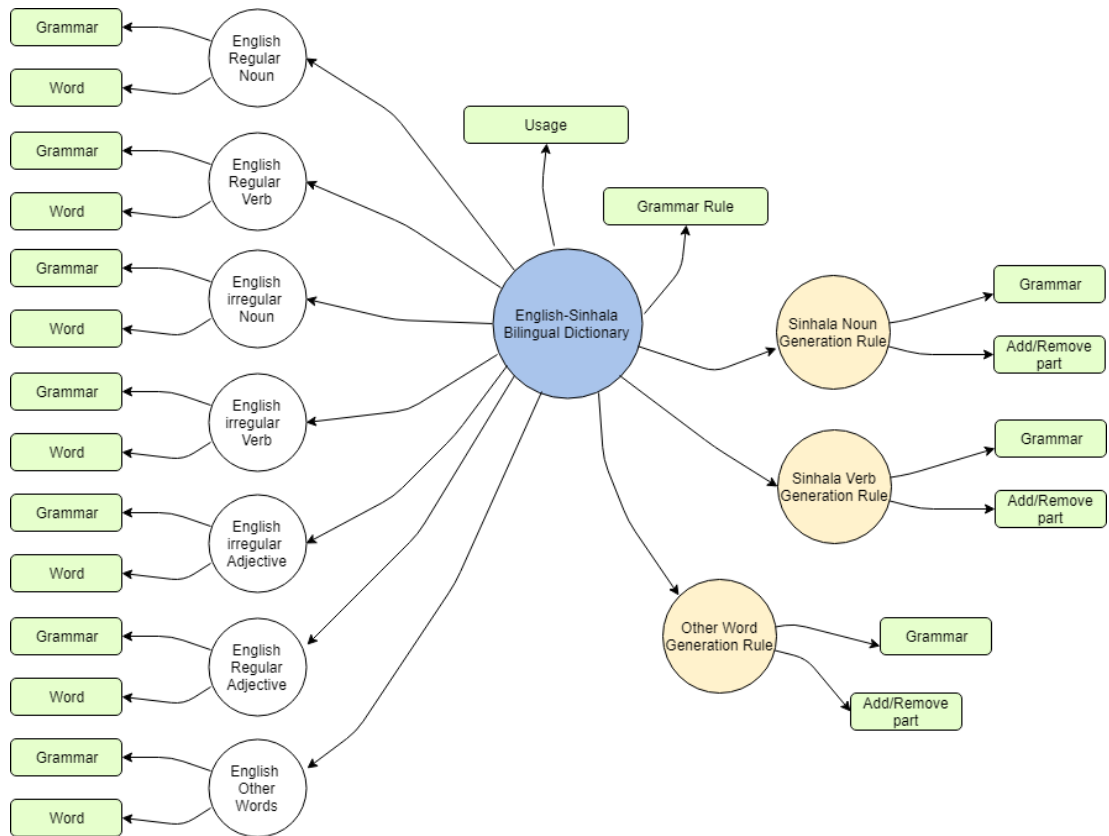


Figure 7.2: Design of the EnSiMaS Ontology

7.2.3 Virtual World

Agents required information is kept in the virtual world. According to the translation requirement, generated morphological, syntactical, semantical and pragmatic information are dynamically stored in the virtual world. For that, it should store morphological information, syntax information of the input sentence(s), semantic information, and Sinhala phrases with their information should also be stored in the virtual world. Figure 7.3 shows the design of EnSiMaS virtual world.

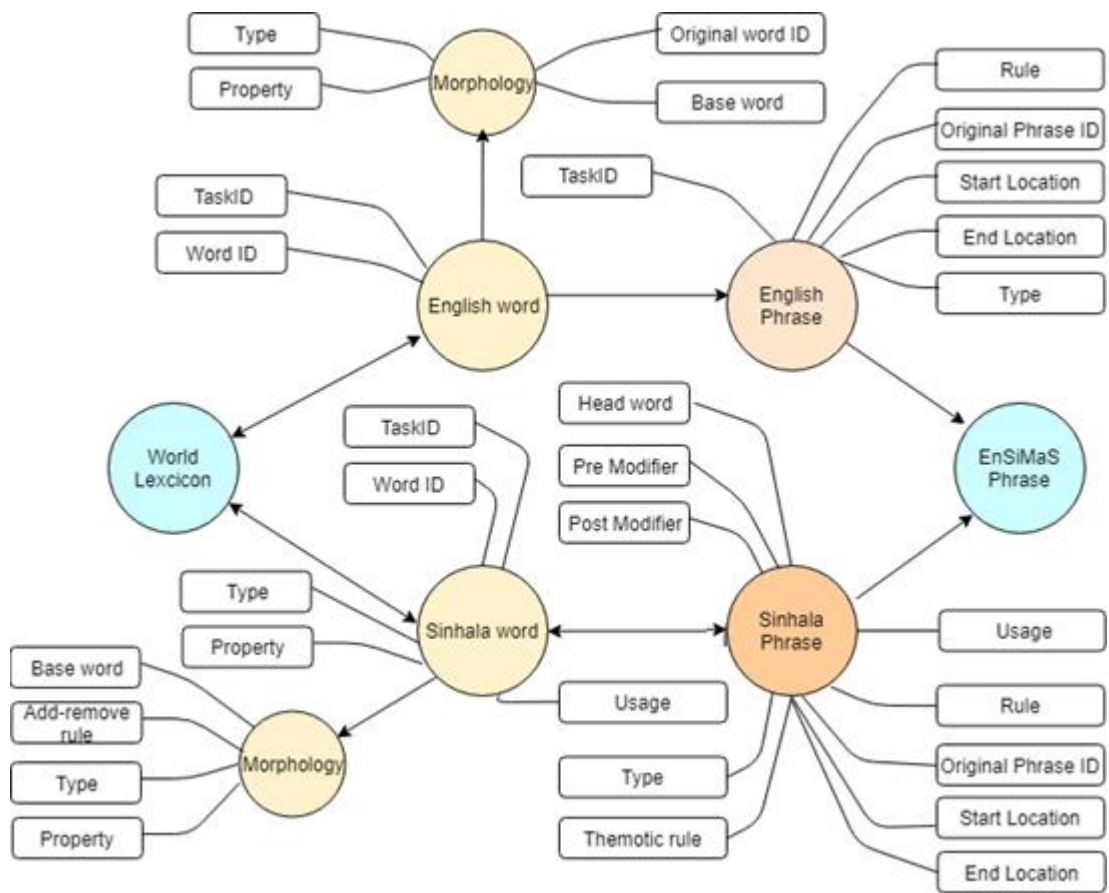


Figure 7.3: Design of the EnSiMaS Virtual world

7.2.4 English Morphological Swarm

The English morphological swarm (EMS) takes the input sentence(s) and conducts morphological analysis for each word in the input. The morphological analyser is used to identify the morphological features of each word. The English morphological analysis swarm has been designed as a swarm of morphological agents in the EnSiMaS system, capable of providing morphological information for the given English words. This swarm comprises of eight morphological processing agents to handle morphology of the English. Figure 7.4 shows the agents' diagram of the English morphological swarm. Note that, each agent handles different types of the English part of speech and provides relevant morphological details.

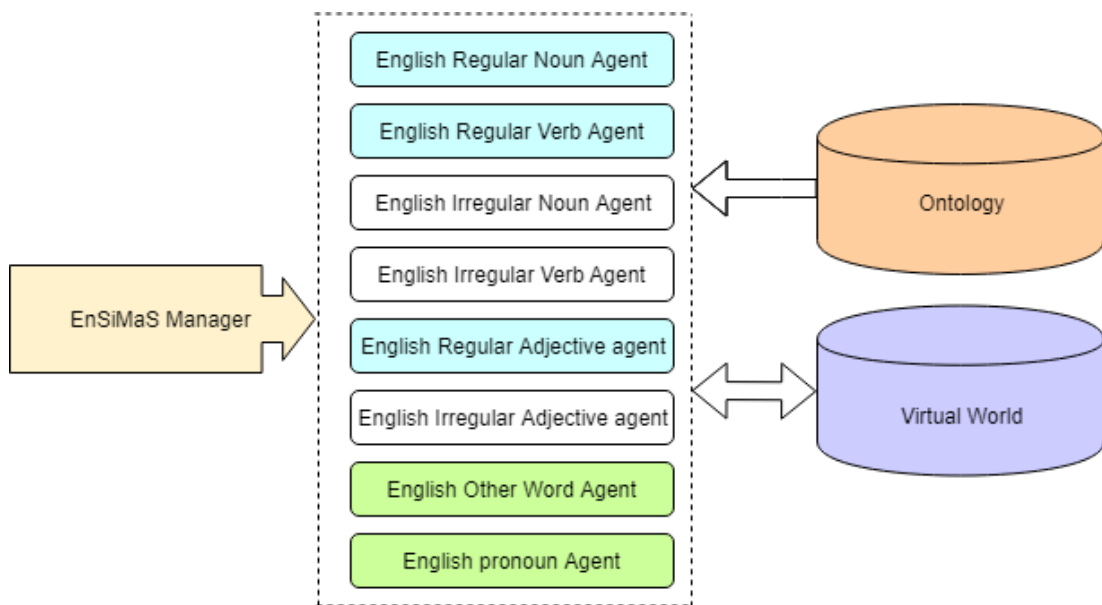


Figure 7.4: Design of the English morphological swarm

7.2.5 English Syntax Analysing Swarm

English syntax-analysing swarm has been designed to analyse syntax (structure of the sentence) of the input sentence(s). This swarm consists of five MaSMT agents, namely a noun phrase search agent, verb phrase search agent, other phrase search agent, syntax identification agent, and thematic relation extraction agent. The English syntax swarm (a swarm of agents) takes the required information from the ontology, including a list of input words (sentence or part of a sentence) and results of the morphological analysis. These phrase search agents take relevant information from the ontology and identify available phrases by scanning in the available tag set that was provided by the morphological swarm. The syntax identification agent should be capable of identifying a suitable structure for the given input English sentence through searching the existing phrases. If the syntax analysis agent is unable to identify the extract structure of the input, it will identify using the relevant phrase agent later in the translation. Somehow, the thematic relation extraction agent uses the identified structure of the sentence and extracts the thematic relationship. After this syntax analysis, each English phrase

consists of phrase structure and relevant thematic relationship for the input. Figure 7.5 shows the agents' diagram of the EnSiMaS syntax-analysing swarm.

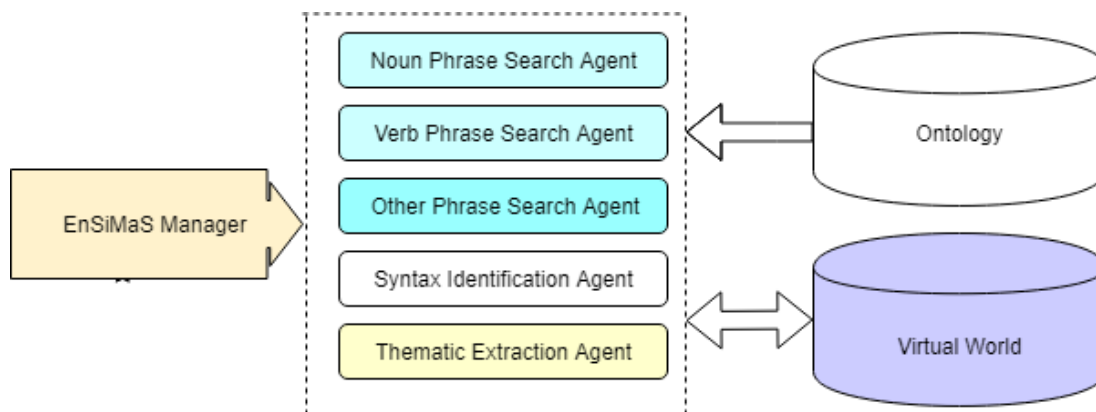


Figure 7.5: Syntax analyzing swarm: agents for English syntax analysis

7.2.6 Bilingual Semantics Swarm

The English-Sinhala semantics swarm has been designed to identify the most suitable Sinhala term for an available English word(s). This swarm also uses an English-Sinhala bilingual dictionary and morphological information on each word to collect appropriate Sinhala terms. All these generated and/or extracted information will be kept in the virtual world for further use by the agents. Note that if a word is not available in the bilingual dictionary, then that word will be recognized as an out-of-vocabulary word. If these out-of-vocabulary words start with a capital letter, then the system will recognise it as a proper noun. Figure 7.6 shows the agents' diagram of the bilingual semantics swarm.

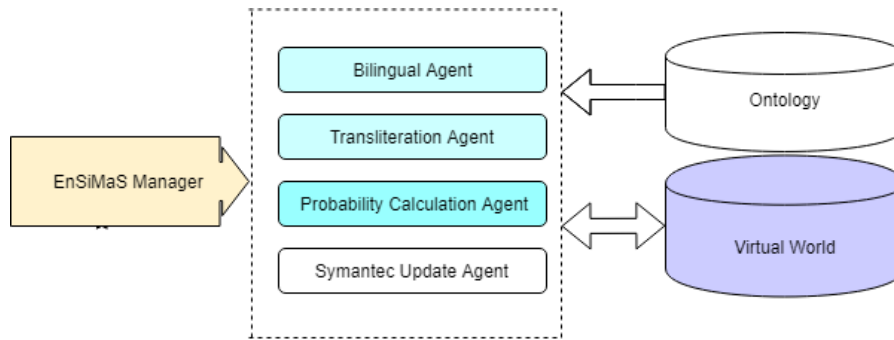


Figure 7.6: Design of the Bilingual Semantics swarm

7.2.7 Sinhala Morphological Swarm

The Sinhala morphological swarm is used to generate relevant Sinhala words for the given root word and the grammar. Note that Sinhala is one of the morphologically rich languages that participates in inflectional and derivational morphology. The purpose of the Sinhala morphological swarm is to generate appropriate forms of a noun or a verb according to the given grammar. Sinhala morphological generation agents take required rules (knowledge to generating an appropriate form of a noun or a verb) from ontology and generate. In here, there are two agents use to generate Sinhala nouns and verbs. Note that other words do not participate word-level conjugation in Sinhala. These Sinhala morphological generation agents use suffix-fitting methods (use the add-remove rule to generate such a word form). The swarm uses the existing suffix-fitting rule set for the Sinhala word generation from the BEES system (“BEES is a rule-based English to Sinhala machine translation system”) [180]. These rules are also stored in the ontology. The Sinhala morphological generation agent reads the relevant information from the ontology and generates appropriate word forms. Finally, generated results are updated into the virtual world for further use. Figure 7.7 shows the Sinhala morphological swarm.

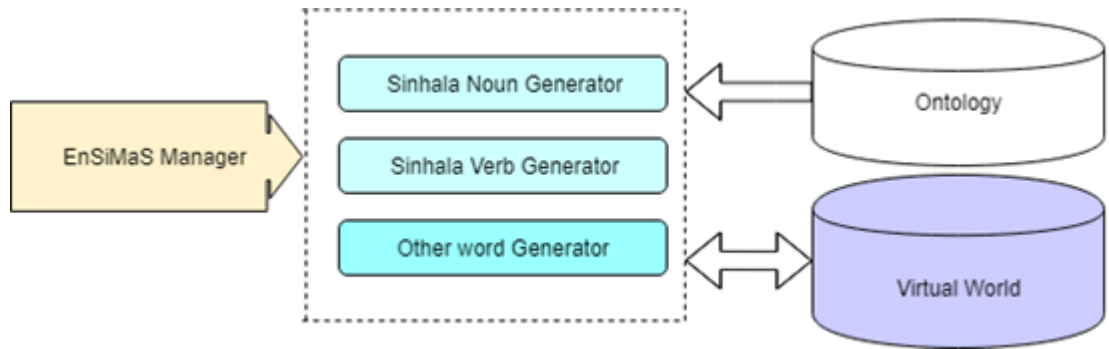


Figure 7.7: Design of the Sinhala morphological swarm

7.2.8 Sinhala Syntactical Swarm

The syntactical swarm capable of re-arranging suitable Sinhala sentence structure for the existing English sentence structure. The Sinhala syntactical swarm consists of an SOV(Subject-Object-Verb) generator agent, a PPOrder (Preposition phrase order) convertor agent, and a complex phrase search agent to use the phrase transfer method to re-generate the Sinhala structure. Note that the structure of the Sinhala sentence and English sentence differs from each other. Therefore, basic transfer rules are available in the ontology and the SOV generation agent to take that information and generate it. As a summary, this agent swarm should be able to re-arrange the order of the existing phrases and make the correct structure of the Sinhala sentence.

7.2.9 Ontological Swarm

All the required lexical resources are available in the ontology. The EnSiMaS agent should be capable of accessing this ontology and taking relevant information. Agent-generated information for the translation (dynamic knowledge) is also stored in the virtual world. Thus, it should require building a virtual word dynamically according to the translation requirements. The main task of the ontology manager agent is to build such information, according to the agent(s) requirements. Thus, this swarm is capable of updating or removing relevant information from the knowledgebase or virtual world.

7.2.10 English Phrase-based Translation Swarm

The English phrase-based translation swarm is the core module of the EnSiMaS system and consists of a dynamically created number of Sinhala phrase agents. Note that each English phrase in the input sentence is mapped into a Sinhala phrase agent that has English phrase information, the number of alternative Sinhala translations for the existing English phrase. According to the structure of the input sentence, existing Sinhala phrase agents can be categorized into subject phrase agent, object phrase agent, verb phrase agent, subject-modifier agent, object-modifier agent etc. Figure 7.8 shows the agents' diagram of the phrase-based translation swarm.

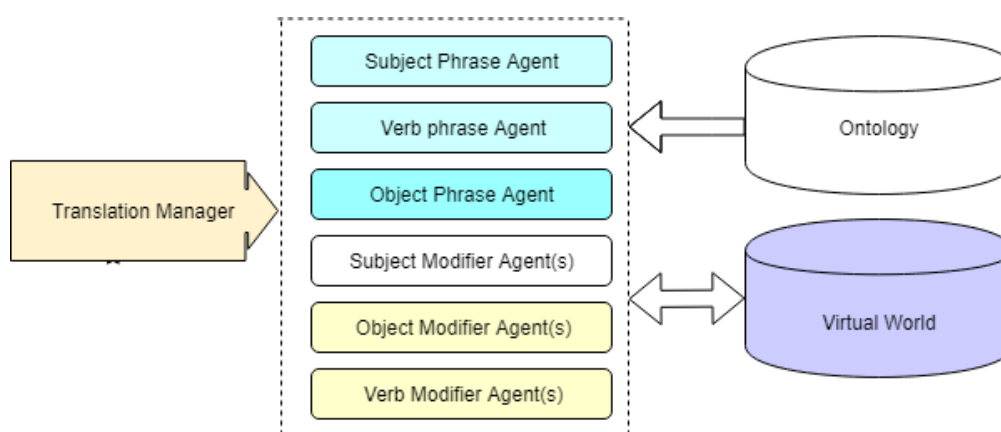


Figure 7.8: Design of the phrase-based translation Swarm

After all Sinhala phrase agents are generated (each English phrase has a Sinhala phrase agent), the translation manager informs the Sinhala phrase agents to start the communication with relevant agents to find suitable solutions. At that time, the Sinhala subject phrase agent and Sinhala verb phrase agent communicate together and select suitable Sinhala translated verbs for the existing Sinhala subjects from the existing Sinhala translations. For that selection, the availability of each alternative solution has been considered (the system should be able to calculate the probability of each Sinhala subject-verb pair), then a suitable Sinhala verb phrase is selected (the Sinhala translation that has maximum probability). With considering the same procedure, the “verb phrase agent” and “object phrase agent” communicate together and take the best solution according to the selected verb phrase for the previously accepted subject

phrase. Note that this phrase selection procedure is based on the concepts behind the psycholinguistic parsing. After subject-verb-object phrase selection, other phrase agents' communicate with relevant head phrase agents and take relevant Sinhala phrases from their alternative Sinhala solutions.

7.3 Summary

This chapter described the design of the EnSiMaS system, which contains 12 modules, namely the GUI, EnSiMaS manager, translation manager, five language processing supporting modules, the ontology manager virtual world and ontology. The next chapter reports the implementation details of the EnSiMaS.

CHAPTER 8

IMPLEMENTATION OF ENSIMAS

8.1 Introduction

The previous chapter reported the design details of the EnSiMaS. The EnSiMaS system consists of 12 modules including GUI, EnSiMaS Manager, Translation manager, five language processing supporting modules, the ontology manager, virtual world and ontology. This chapter gives a brief description of how each module and the tools have been implemented.

8.2 EnSiMaS Ontology

The EnSiMaS ontology consists of all the required morphological, syntactical and semantic information on both languages. English Language resources and morphological rules are also stored in the English dictionary. The English dictionary consists of eight tables to store pronouns, regular nouns, regular verbs, regular adjectives, irregular verbs, irregular verb and other words are also stored in a separate table. Table 8.1 shows the summary of the English dictionary with records available on each table.

Table 8.1: Statistics of the EnSiMaS Knowledgebase

English Part of speech category	Number of Records
Regular Noun	12320
Irregular Noun	9860
Pronoun	53
Regular Verb	2258
irregular Verb	8459
Regular Adjective	7722

Other words (Adverb, irregular adjectives, conjunctions and articles)	1121
English-Sinhala Bilingual Dictionary	87386
English Morphological rules	19
English Morphological rules	165
Phrase rules	78

Note that, sample data set for each table, the structure of each table and sample SQL code is listed below.

8.2.1 English Pronoun Table

The pronoun table consists of English pronouns and the morphological details of each pronoun. Figure 8.1 shows the structure and selected sample data available on the pronoun table.

id	word_id	prop	word
1	80000001	nun-com-fpe-sin-sub	i
2	80000002	nun-com-fpe-sin-obj	me
4	80000003	nun-com-fpe-sin-ppn	mine
5	80000004	nun-com-fpe-sin-ref	myself
6	80000005	nun-com-fpe-plu-sub	we
7	80000006	nun-com-fpe-plu-obj	us
8	80000007	nun-com-fpe-plu-pde	our
9	80000008	nun-com-fpe-plu-ref	ourselves
10	80000009	nun-com-fpe-sin-pde	my
12	80000010	nun-com-fpe-sin-ppr	ours
13	80000011	nun-com-spe-sin-sub	you

Name	Type	Not	PK	AI
id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
word_id	integer	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
prop	varchar (20)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
word	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>


```
CREATE TABLE `eng_pro_noun` (
  `id` integer PRIMARY KEY AUTOINCREMENT,
  `word_id` integer NOT NULL,
  `prop` varchar ( 20 ) NOT NULL,
  `word` varchar ( 50 ) NOT NULL
);
```

Figure 8.1: Structure and sample data on the pronoun table

8.2.2 English Regular Noun table

The English Regular noun table consists of regular English nouns and the morphological details of each regular noun. Figure 8.2 shows the structure and selected sample data available on the regular noun table.

word_id	prop	word
20000045	nun-tpe-nue	abstention
20000046	nun-tpe-nue	abstinence
20000047	nun-tpe-nue	abstract
20000048	nun-tpe-nue	abstraction
20000049	nun-tpe-nue	abundance
20000050	nun-tpe-nue	abuse
20000051	nun-tpe-nue	abutment
20000052	nun-tpe-nue	abutter
20000053	nun-tpe-nue	abysm

Name	Type	Not	PK	AI
word_id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
prop	varchar (20)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
word	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>


```
CREATE TABLE `enrgnun` (
  `word_id` integer PRIMARY KEY AUTOINCREMENT,
  `prop` varchar ( 20 ) NOT NULL,
  `word` varchar ( 50 ) NOT NULL
);
```

Figure 8.2: Structure and sample data on the regular noun table

8.2.3 English Regular Verb Table

This table consists of the English regular verbs and the morphological details of the regular verb. Figure 8.3 shows the selected sample data and the structure of the regular verb table.

word_id	word	prop
40001834	smarten	veb
40001835	smash	veb
40001836	smatter	veb
40001837	smear	veb
40001838	smelt	veb
40001839	smirch	veb
40001840	smirk	veb

Name	Type	Not	PK	AI
word_id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
word	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
prop	varchar (20)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>


```
CREATE TABLE `enregvb` (
  `word_id` integer PRIMARY KEY AUTOINCREMENT,
  `word` varchar ( 50 ) NOT NULL,
  `prop` varchar ( 20 ) NOT NULL
);
```

Figure 8.3: Structure and sample data on the regular verb

8.2.4 English Irregular noun table

The English irregular noun table consists of irregular English nouns and the morphological details of the regular noun. Figure 8.4 shows the selected sample data and structure of the irregular noun table

id	word_id	prop	word
2	10000001	nun-tpe-nue-plu	a-bombs
3	10000002	nun-tpe-nue-sig	abacus
4	10000002	nun-tpe-nue-plu	abacuses
7	10000004	nun-tpe-nue-sig	abandon
9	10000005	nun-tpe-nue-sig	abandonment
10	10000005	nun-tpe-nue-plu	abandonments
15	10000008	nun-tpe-nue-sig	abbacy
16	10000008	nun-tpe-nue-plu	abbacies
17	10000009	nun-tpe-nue-sig	abbess
18	10000009	nun-tpe-nue-plu	abbesses

Name	Type	Not	PK	AI
id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
word_id	integer	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
prop	varchar (20)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
word	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

```

CREATE TABLE `enirnun` (
  `id` integer PRIMARY KEY AUTOINCREMENT,
  `word_id` integer NOT NULL,
  `prop` varchar ( 20 ) NOT NULL,
  `word` varchar ( 50 ) NOT NULL
);

```

Figure 8.4: Structure and sample data on the irregular noun

8.2.5 English Irregular verb table

This table consists of irregular English verbs and the morphological details of the irregular verb. Figure 8.5 shows the selected sample data and structure of the irregular verb table.

id	word_id	prop	word
8882	30002961	veb-inf	write
8883	30002961	veb-pat	wrote
8884	30002961	veb-pap	written
8885	30002962	veb-inf	writhed
8886	30002962	veb-pat	writhed
8887	30002962	veb-pap	writhed
8891	30002964	veb-inf	yacht
8892	30002964	veb-pat	yachted
8893	30002964	veb-pap	yachted

Name	Type	Not	PK	AI
id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
word_id	integer	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
prop	varchar (20)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
word	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

```

CREATE TABLE `enirrvb` (
  `id` integer PRIMARY KEY AUTOINCREMENT,
  `word_id` integer NOT NULL,
  `prop` varchar ( 20 ) NOT NULL,
  `word` varchar ( 50 ) NOT NULL
);

```

Figure 8.5: Structure and sample data on the irregular Verb table

8.2.6 English Regular Adjective Table

The English regular adjective table consists of regular English adjectives and the morphological details of the regular adjectives. Figure 8.6 shows the selected sample data and the structure of the regular adjective table

word_id	prop	word
70005795	adj	shaven
70005797	adj	sheeny
70005798	adj	sheepish
70005799	adj	sheer
70005800	adj	shelled
70005801	adj	shelter
70005802	adj	shelving

Name	Type	Not	PK	AI
word_id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
prop	varchar (25)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
word	varchar (25)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>


```
CREATE TABLE `enrgadj` (
  `word_id` integer PRIMARY KEY AUTOINCREMENT,
  `prop` varchar ( 25 ) NOT NULL,
  `word` varchar ( 25 ) NOT NULL
);
```

Figure 8.6: Structure and sample data on the regular adjective table

8.2.7 Other word table

Other words table consists of the adverbs, conjunctions, other parts of speech and the morphological details of each word. Figure 8.7 shows the selected sample data and the structure of the other word table.

word_id	prop	word
60000027	adv	accusingly
60000028	adv	across
60000029	prp	across
60000030	adv	activate
60000031	adv	actively
60000032	adv	actually
60000033	adv	acutely
60000034	adv	ad-infinitum
60000035	adv	adagio

Name	Type	Not	PK	AI
word_id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
prop	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
word	varchar (25)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>


```
CREATE TABLE `enirrd` (
  `word_id` integer PRIMARY KEY AUTOINCREMENT,
  `prop` varchar ( 50 ) NOT NULL,
  `word` varchar ( 25 ) NOT NULL
);
```

Figure 8.7: Structure and sample data on the other word table

8.2.8 English-Sinhala Bilingual Dictionary

This dictionary consists of English and Sinhala words with the Google search index and Human Usage Index. Google Search Index value take through the Google search result, and human-usage value take form EnSiMaS Dictionary. Figure 8.8 shows the structure of the bilingual dictionary and a few sample records.

	id	engword	sinword	engid	sinid	type	us	ruleid	hus
1	1	a-bomb	පර් බොම්බ පවිත්රිමය	10000001	0	nun	163	101	0
2	2	aback	පැපට	60000001	0	adv	184000	101	0
3	3	aback	පැප්පෙප්	60000001	0	adv	278000	101	0
4	4	aback	පසු පසට	60000001	0	adv	65900	101	0
5	5	aback	පසුපසට	60000001	0	adv	206000	101	0
6	6	aback	පිටපසට	60000001	0	adv	30500	101	0
7	7	abacus	පසුපසු පිස	10000002	0	nun	1540	101	0
8	8	abacus	පසුපසු පිසුරු	10000002	0	nun	107	101	0
9	9	abacus	පසුපසු පිසුරු	10000002	0	nun	1170	101	0
10	10	abaft	අවරට	60000002	0	adv	36900	101	0
11	11	abaft	පැවරින පසු පසුපසට	60000002	0	adv	312	101	0

Name	Type	Not	PK	AI	U	Default
id	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
engword	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
sinword	varchar (50)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
engid	integer	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
sinid	integer	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
type	varchar (10)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
us	integer	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
ruleid	integer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	101
hus	INTEGER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0

Figure 8.8: Structure and sample data on the bilingual dictionary

8.2.9 Morphological rules for English words

The English morphological analysis system uses the suffix-remove method to identify the headword (Root word). Thus, all the required add-remove rules are defined in the morphological rules table. Figure 8.9 shows the rules, which are defined to identify English morphology.

id	table_name	add_rule	rem_rule	prop_value	rule_type
1	eng_reg_verb		s	spr	find
2	eng_reg_verb		ed	pap	find
3	eng_reg_noun			sin	get
4	eng_reg_verb				get
5	eng_reg_noun		s	plu	find
6	eng_reg_verb		ing	ppa	find
7	eng_irr_verb				get
8	eng_irr_noun				get
9	eng_reg_noun		`s	sin-spo	find
10	eng_irr_word				get
11	eng_reg_adjt				get
12	eng_reg_noun		s`	plu-plp	find
13	eng_reg_verb		ed	pat	find
14	eng_irr_verb		ing	ppa	rule01
15	eng_irr_verb		s	spr	rule01
16	eng_pro_noun				get
17	eng_reg_adjt		er	com	find
18	eng_reg_adjt		est	sup	find
19	eng_irr_adjt				get

Figure 8.9: Sample data for English morphological rules

8.2.10 Syntax Rule for English phrases

The English syntax analyzing system (Phrase-based chunker) requires knowledge of English phrases. Therefore, syntax rules are defined with considering the existing phrase structure (noun phrase, verb phrase, etc.) for English. Figure 8.10 shows some existing rules with their information.

id	morp	count	prop	tp	gp	rwid	rmop
44	VBZ	1	ACT-SPT	VP	0	1	VBZ
45	VBP	1	ACT-SPT	VP	0	1	VBP
46	VBD	1	ACT-PAT	VP	0	1	VBD
47	XBF,XGG,PTO,VBP	4	ACT-FUT	VP	0	4	VBP
48	XBF,VBG	2	ACT-PRC	VP	0	2	VBG
49	XHX,VBN	2	ACT-PRP	VP	0	2	VBN
50	XHS,XBN,VBG	3	ACT-PRPC	VP	0	3	VBG
51	XWS,VBG	2	ACT-PAC	VP	0	2	VBG
52	XWE,VBG	2	ACT-PAC	VP	0	2	VBG
53	XHV,XBN,VBG	3	ACT-PRPC	VP	0	3	VBG
54	XHD,VBN	2	ACT-PAP	VP	0	2	VBN
55	XHD,XBN,VBG	3	ACT-PAPC	VP	0	3	VBG
56	XSH,VBP	2	ACT-FUT	VP	0	2	VBP
57	XSH,XBE,VBG	3	ACT-FUC	VP	0	3	VBG
58	XSH,XHV,VBN	3	ACT-FUP	VP	0	3	VBN
59	XSH,XHV,XBN,VBG	4	ACT-FUPC	VP	0	4	VBG
60	XWL,XHV,XBN,VBG	4	ACT-FUPC	VP	0	4	VBG
61	RBX,VBZ	2	ACT-SPT	VP	0	2	VBZ
62	FPS	1	FPS	NP	0	1	FPS
63	FPP	1	FPP	NP	0	1	FPP
64	SPS	1	SPS	NP	0	1	SPS
65	SPP	1	SPP	NP	0	1	SPP

Name	Type	Not	PK	AI
id	INTEGER	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
morp	TEXT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
count	INTEGER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
prop	TEXT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
tp	TEXT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
gp	INTEGER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
rwid	INTEGER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
rmop	TEXT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ex	TEXT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>


```

CREATE TABLE `NPSyntaxOnto` (
  `id` INTEGER PRIMARY KEY AUTOINCREMENT,
  `morp` TEXT,
  `count` INTEGER DEFAULT 0,
  `prop` TEXT,
  `tp` TEXT,
  `gp` INTEGER DEFAULT 0,
  `rwid` INTEGER,
  `rmop` TEXT,
  `ex` TEXT
);

```

Figure 8.10: Selected rules to detect English phrases (Phrase rules)

8.3 EnSiMaS Virtual World

“Virtual World” is represented as the current environment of the multi-agent system. The EnSiMaS system dynamically generates the required knowledge through the existing ontology and input sentence(s). This section briefly explains the language model, which was used to create the virtual world. For the translation purpose, the following language model was used.

1. **English Word:** This model consists of information on given input words, including the word and its morphology.
2. **English Morphology:** This model stores morphological information. After the English morphological analysis, the English morphology is identified for each English word.
3. **English Phrase:** After the syntax analysis, English phrases are identified for the existing English sentence(s). These phrases’ information is stored in an English phrase.

4. **Sinhala Phrase:** Generated Sinhala phrases are stored in a Sinhala phrase object.
5. **Word Lexicon:** According to the English-Sinhala bilingual results, English words have zero or more Sinhala terms. The Word Lexicon model is used to store that information.
6. **EnSiMaS Phrase:** An English phrase can be translated into a number of Sinhala phrases. All this information is stored in an EnSiMaS phrase.
7. **EnSiMaS Phrase list:** Translation results are stored in a list of EnSiMaS phrases. Note that the EnSiMaS phrase list consists of target language sentences.

The design of each language model has been described below.

8.3.1 The English Word and Word List

The model “EnglishWord” is the primary class of the EnSiMaS system. Figure 8.11 shows the class diagram of the “EnglishWord”, which is used to store English word information. The English word object consists of a translation task ID, word ID (position of the sentence), word and its morphology. Note that the field morphology has been added after the morphological analysis.

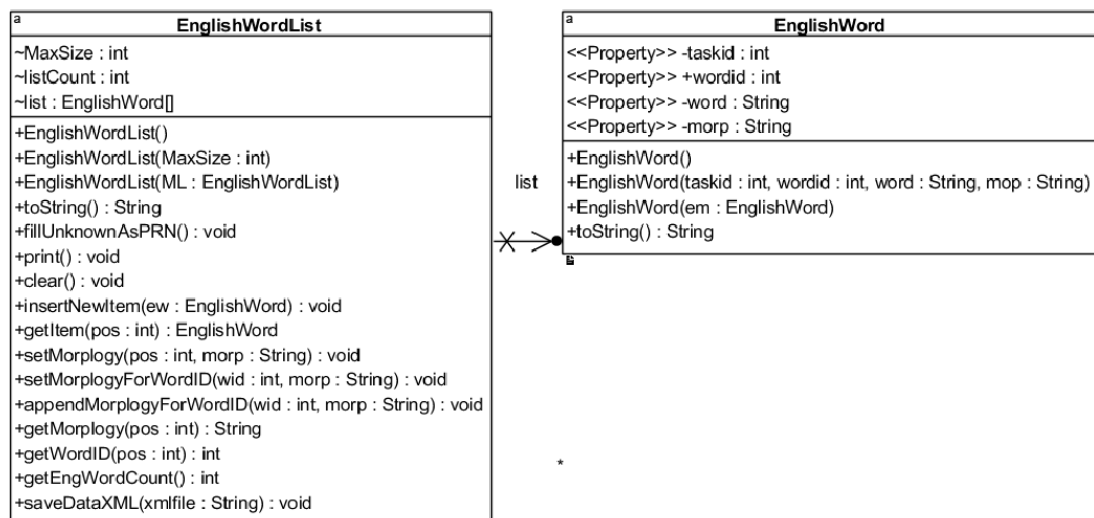


Figure 8.11: Class diagrams of the English word and English wordlist

The English Word Morphology

English morphological information has been updated by the English morphological analyzing system. Figure 8.12 shows a class diagram of the English word morphology. This class consists of the word ID and the existing English word. In addition to that, it takes the original base word of the existing word. For instance, the word “playing” has the base word “play”.

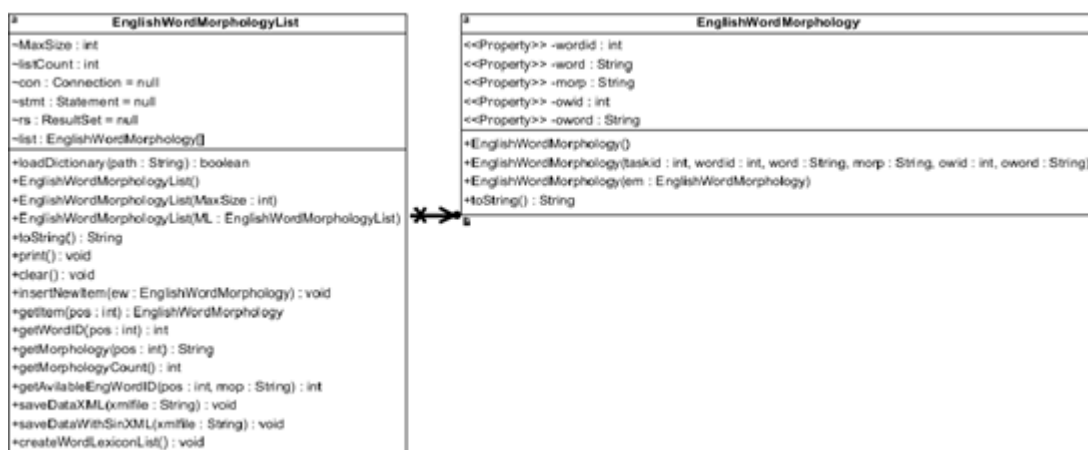


Figure 8.12: Class diagrams of the English word morphology and morphology list

8.3.2 The English Phrase and Phrase List

The English phrase consists of one or more English word. The English phrase class is used to store phrase information on the EnSiMaS system. Figure 8.13 shows the English phrase class. The English phrase object consists of relevant information on the English phrase, including the starting point, endpoint, syntax rules, etc.

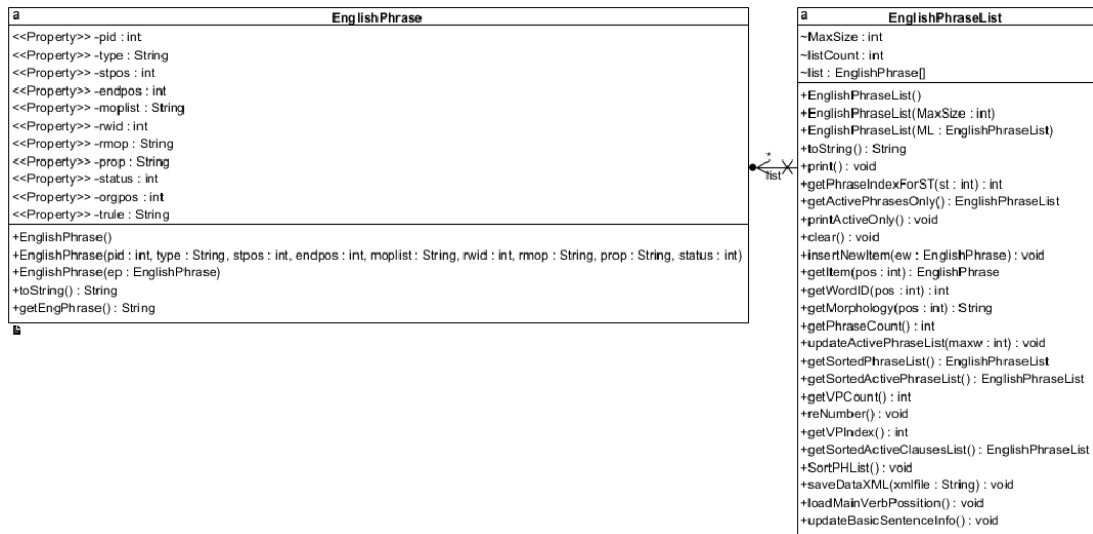


Figure 8.13: Class diagrams of the English phrase and English phrase list

8.3.3 The Sinhala Phrase and Sinhala phrase list

The translated Sinhala phrase is stored in a Sinhala phrase object. Figure 8.14 shows the class diagram of the Sinhala Phrase. The Sinhala phrase consists of the phrase type, probability, original Sinhala headword, morphologically generated Sinhala headword, and morphological information for each word. The pre-modifier and post-modifier for the English phrase are also recorded.

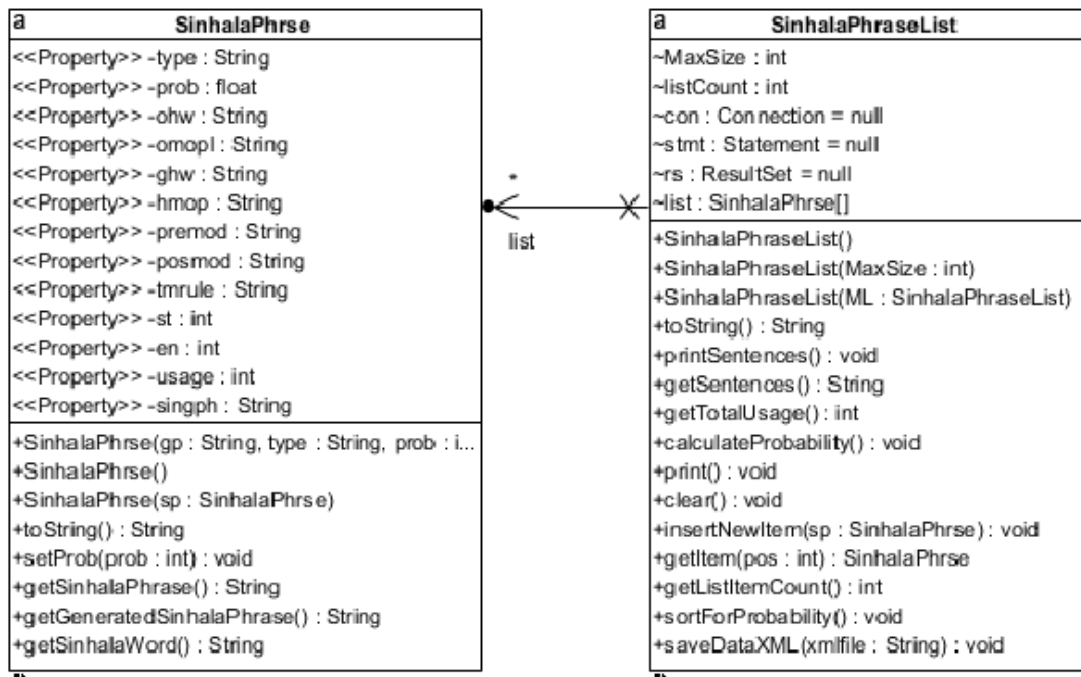


Figure 8.14: Class diagrams of the Sinhala Phrase and Sinhala phrase list

8.3.4 The Sinhala word Lexicon

The English-Sinhala bilingual dictionary consists of relevant Sinhala terms for existing English words or phrases. In addition to the Sinhala term, it also consists of Google search usage, human selection index, and morphology for the existing English word. All this information is stored in a lexicon object. Figure 8.15 shows the class diagram of the Sinhala word lexicon and Sinhala word lexicon list.

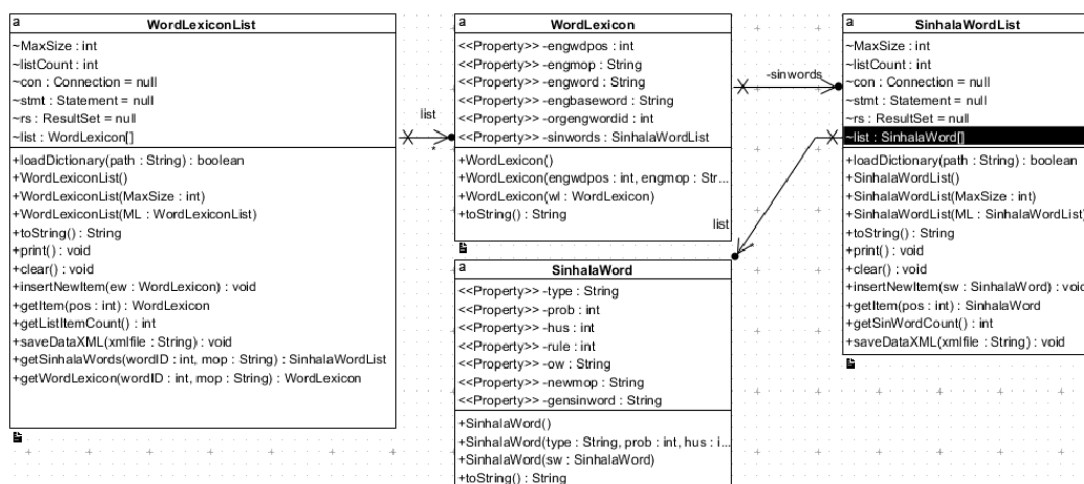


Figure 8.15: Class diagrams of the Sinhala word lexicon and Sinhala word lexicon list of the EnSiMaS

8.3.5 The EnSiMaS Phrase

According to the approach discussed in this thesis, the English phrase is translated into a number of Sinhala phrases. It should be required to store all the above information in the virtual world. EnSiMaS phrase has been designed to store all phrases information available in the translation (generated dynamic information stored in the virtual world). Figure 8.16 shows the structure of the EnSiMaS phrase. Note that the EnSiMaS phrase consists of the Sinhala phrase list, the English phrase, and other relevant information such as the best phrase, top phrase, and the existing number of phrase and thematic rules for the existing English phrase.

The Sinhala phrase list: Consists of some translated Sinhala phrases for the given English phrase.

Best Phrase: After the Sinhala phrase agents' communication, (subject-verb or object-verb) the Sinhala phrase, with the maximum probability, is moving to the best phrase (if the phrase is a verb phrase, the probability take with considering the Sinhala subject phrase and verb phrase).

Top Phrase: Considering the existing Sinhala translations (The Sinhala phrase on the EnSiMaS phrase list), the phrase with the maximum probability set as the top phrase. (that takes only considering the phrases on the Sinhala phrase list)

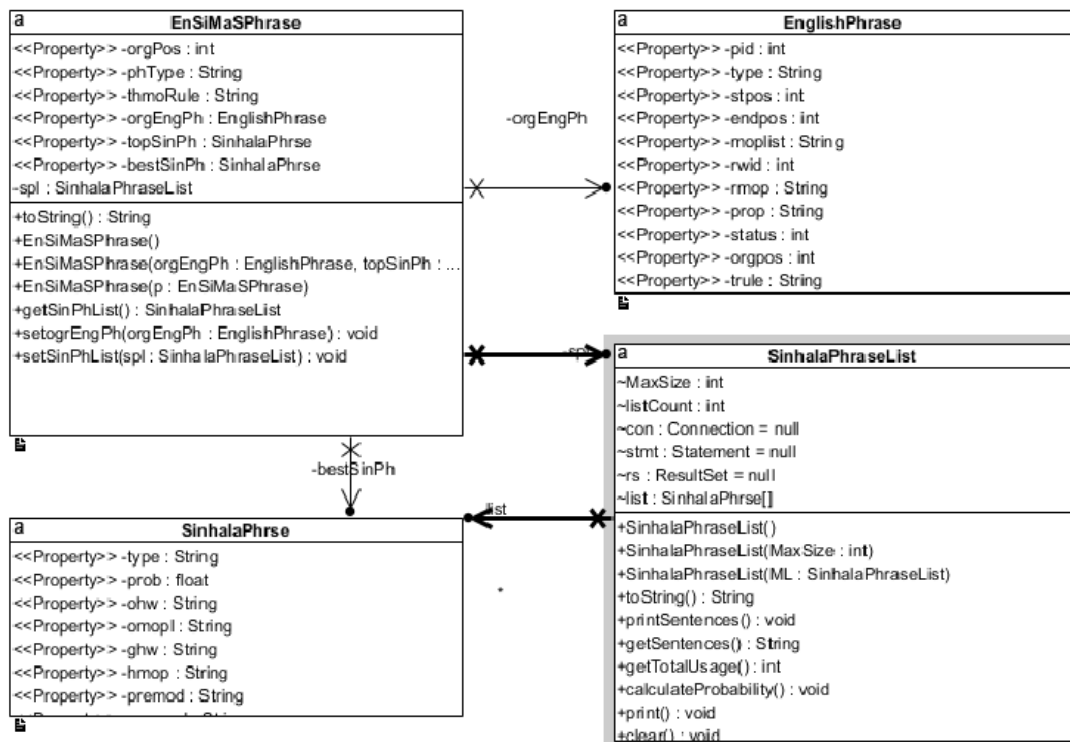


Figure 8.16: Class diagram of the EnSiMaS phrase

8.3.6 The EnSiMaS Phrase List

The Sinhala sentence can be considered as one or more Sinhala phrases which are listed some syntactical order. The EnSiMaS phrase list consists of one or more EnSiMaS phrases that represent a sentence. Figure 8.17 shows the structure of the EnSiMaS phrase list.

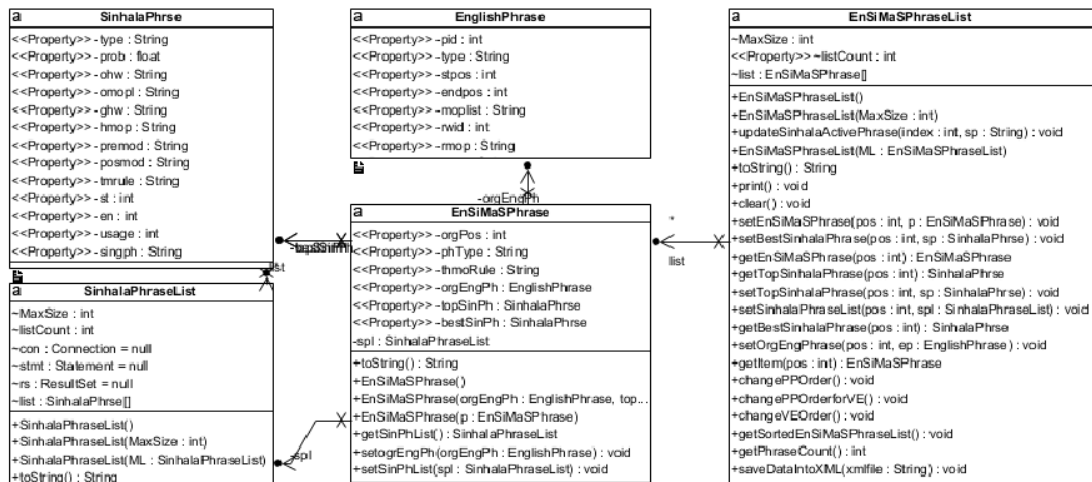


Figure 8.17: Class diagram of the EnSiMaS phrase list

8.3.7 The EnSiMaS Sentence info

After generating correspondent Sinhala translations for the given English sentence, Sinhala sentence information is stored into EnSiMaS sentence info object. This class object also consists of thematic information for the translated Sinhala sentence. The information available on the EnSiMaS sentence info is used to identify pragmatic analysis or further required.

Up to this point, Ontology model has been discussed. The rest of the section briefly explains the implementation of the six-language processing model for language translation. Also, the implementation of the EnSiMaS will be discussed below.

8.4 EnSiMaS Agents

This section briefly describes implementation details of the EnSiMaS system that consists of seven language processing systems, Sinhala phrase agents and two controlling agents, namely EnSiMaS Manager and Translation Manager.

8.4.1 EnSiMaS Manager Agent

EnSiMaS manager agent (MaSMT controller) is the key control agent of the system that control all the other agents, including the translation manager and language

processing agents. The manager (Agent controller of the MaSMT) agent can directly control several MaSMT agents. Thus, the EnSiMaS manager consists of seven language processing swarms for language processing, and the translation manager is used for handling the translation process. This EnSiMaS manager reads the input sentence and provides translated outputs to the GUI.

8.4.2 English Morphological System

The English morphological swarm has been implemented using MaSMT agents. Table 8.2 shows the implantation details of each agent in the morphological swarm.

Table 8.2: Agents of the English morphological swarm

No	Agent Name	Agent ID	Task
1	EnglishIrrNounMopAgent	analysis.101@ema	Identify irregular noun
2	EnglishRegVerbMopAgent	analysis.102@ema	Identify regular verb
3	EnglishRegNounMopAgent	analysis.103@ema	Identify regular verb
4	EnglishProNounMopAgent	analysis.104@ema	Identify Pronoun
5	EnglishIrrWordMopAgent	analysis.105@ema	Identify irregular Other words+ Auxiliary verbs
6	EnglishIrrVerbMopAgent	analysis.106@ema	Identify irregular Verb
7	EnglishIrrAdjMopAgent	analysis.107@ema	Identify irregular adjective
8	EnglishRegAdjMopAgent	analysis.108@ema	Identify regular adjective

These morphological agents use the suffix-removing/root-fixing method to identify existing English morphology. According to the root-fixing approach, add-remove rules are available for the English word. Each agent takes its rules and relevant morphological information from the knowledge base and waits for messages to do the analysis process. The EnSiMaS manager sends the message “analyze-eng-morp” with the input sentence, then the agent reads the input and analyses the morphology. The

result has been written in the virtual world for further use. Table 8.3 shows the morphological tag set which is used to identify words. Then, the agent informs the completed task activity for the EnSiMaS Manager. According to the diagram (Figure 8.18), the English regular noun agent should be capable of identifying English regular nouns from the existing English source. The English regular verb can identify by the regular verb agent. Similarly, all the above-related part of speech tags are identified. After completing the task, the agent informs it into its manager and waiting for the next task. In general, the morphological agent takes relevant morphological rules from the ontology and waits for messages. If a message is available, then do the action according to that. Figure 8.18 shows the activity diagram for the morphological agent.

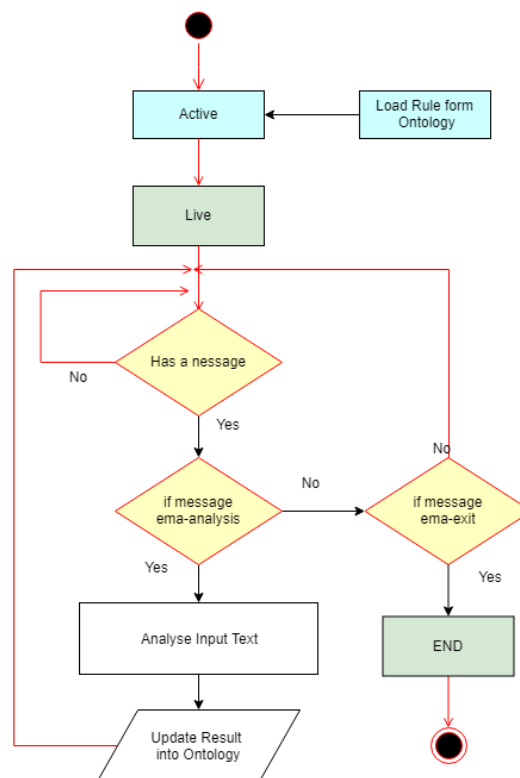


Figure 8.18: Activity diagram of the morphological agent

Figure 8.19 shows the agent communication diagram of the morphological swarm. According to the communication diagram, EnSiMaS manager broadcast messages to its client agent (Morphological processing agents only) then each agent take the message and work according to that. If Morphological analyzing process complete, then send a reply message to EnSiMaS manager.

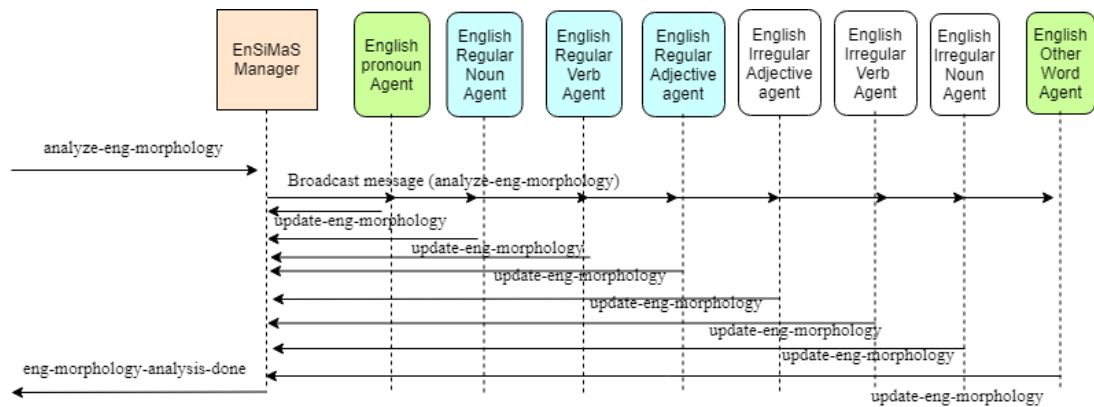


Figure 8.19: Communication diagram of the EMS

Table 8.3: Morphological TAGs

No	TAG	Description
1	VBP	Infinitive
2	VBZ	Simple Present
3	VBD	Past
4	VBN	Past Participle
5	VBG	Present Participle
6	NWS	Singular Noun
7	NWP	Plural
8	NPS	Singular Possessive
9	NPP	Plural Possessive
10	PNS	Subject Pronouns
11	PNO	Object Pronouns
12	PNA	Possessive Pronouns
13	PNR	Reflexive Pronouns
14	JJX	Adjective
15	JJC	Adjective (Comparative)
16	JJS	Adjective(Superlative)

17	CON	Conjunction
18	PRP	Preposition
19	DET	Articles
20	RBX	Adverb

8.4.3 English Syntax Swarm

The English syntax swarm is used to identify the syntax structure of the given sentence. For that, the English syntax swarm uses the result of the morphological analysis with an input sentence. According to the design of the EnSiMaS, each agent has been designed as a MaSMT agent. Table 8.4 shows the implantation details of each agent in the English syntactical swarm.

Table 8.4: Agents' details of the English syntax swarm

No	Agent Name	Agent ID	Task
1	EnglishNounPhraseSearchAgent	analysis.201@esa	Investigate the possible Noun phrases
2	EnglishVerbPhraseSearchAgent	analysis.202@esa	Investigate the possible Verb phrases
3	EnglishOtherPhraseSearchAgent	analysis.203@esa	Investigate the possible Other phrases
4	SyntaxIdentificationAgent	analysis.204@ema	If possible identify suitable syntax form the existing phrases
5	ThemoticRelationExtraction Agent	analysis.205@ema	Considers the syntax and identify thematic relation in between each phrase

Note that all the required rules are available in the knowledge base to search available phrases. Thus, search agents (noun phrase, verb phrase, and other phrases) take relevant information from the ontology and search through the existing

morphologies. After phrase investigation, the syntax identification agent searches through the available phrases and makes correct syntax structure for the given sentence. After syntax identification, the thematic relation extraction agent searches the relation in between the subject-verb and object-verb. Finally, the thematic relation extraction agent sends the notification message to the EnSiMaS Manager to inform that the task is complete. Figure 8.20 shows the general activity diagram for the phrase search agents and Figure 8.21 shows the communication diagram of each agent.

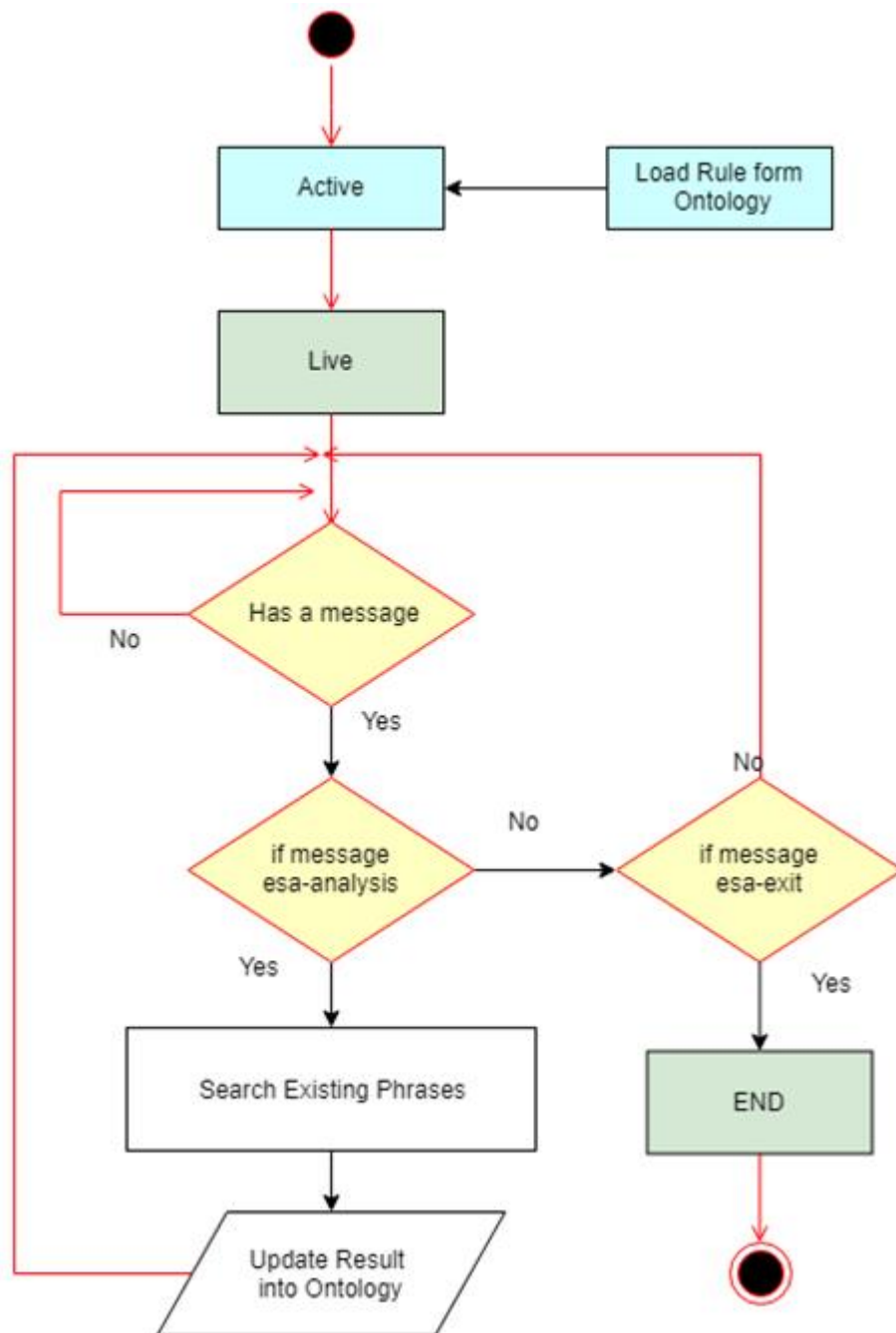


Figure 8.20: Activity diagram of the English Syntax Swarm

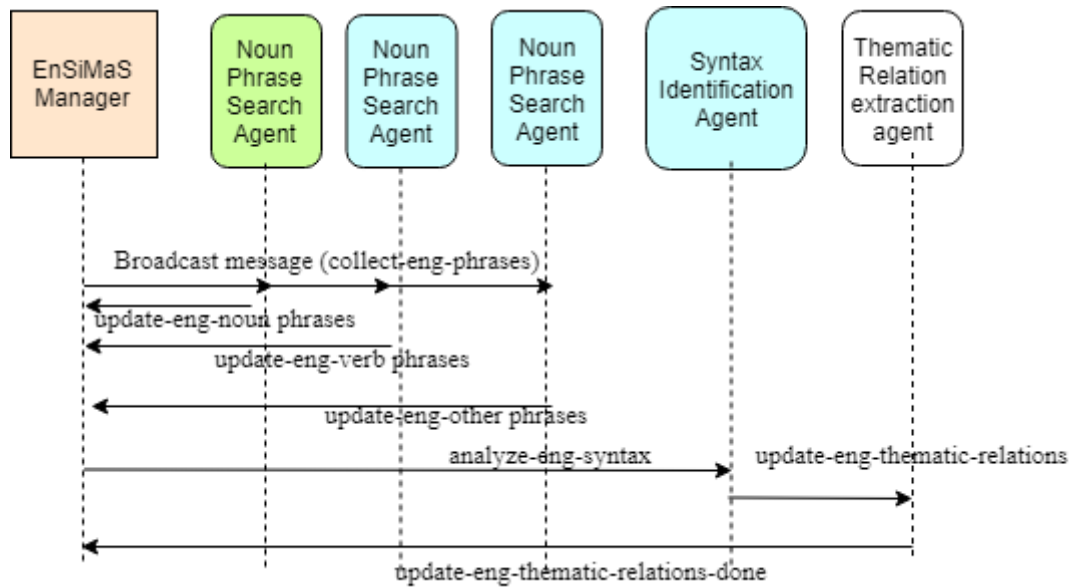


Figure 8.21: Communication diagram of the syntactical swarm

8.4.4 Bilingual Semantic Swarm

Relevant Sinhala terms have been identified through the Bilingual Semantics Swarm (BSS). This BSS comprises of four agents, namely the bilingual agent, transliteration agent, probability calculation agent, and the semantics update agent. Table 8.5 shows the implementation details of the Bilingual semantics swarm. After the morphological analysis, the EnSiMaS manager sends a message to the bilingual agent to “update-Sinhala-ontology”, then the agent takes the morphology list from the virtual world and available Sinhala terms for each English word. If the Sinhala term is not available, it sends a message to the transliteration agent to transliterate the English term. After collecting all the Sinhala terms, the bilingual agent sends a message to the probability calculation agent to calculate the available probability for each word. Then all the generated information is updated into the virtual world through the semantics update agent. The agent communication diagram of the Sinhala semantic swarm is given in Figure 8.22.

Table 8.5: Agents' details of the Bilingual Semantic swarm

No	Agent Name	Agent ID	Task
1	BilingualAgent	analysis.301@bss	Search available Sinhala term from the bilingual dictionary
2	TransliterationAgent	analysis.302@bss	If Sinhala term is not available in the bilingual dictionary, then transliterate it
3	probabilityCalculationAgent	analysis.303@bss	Calculate the probability for a particular Sinhala term compare with another term on the same English word
4	semantics update agent	analysis.304@bss	All the ontology update process handle by this agent

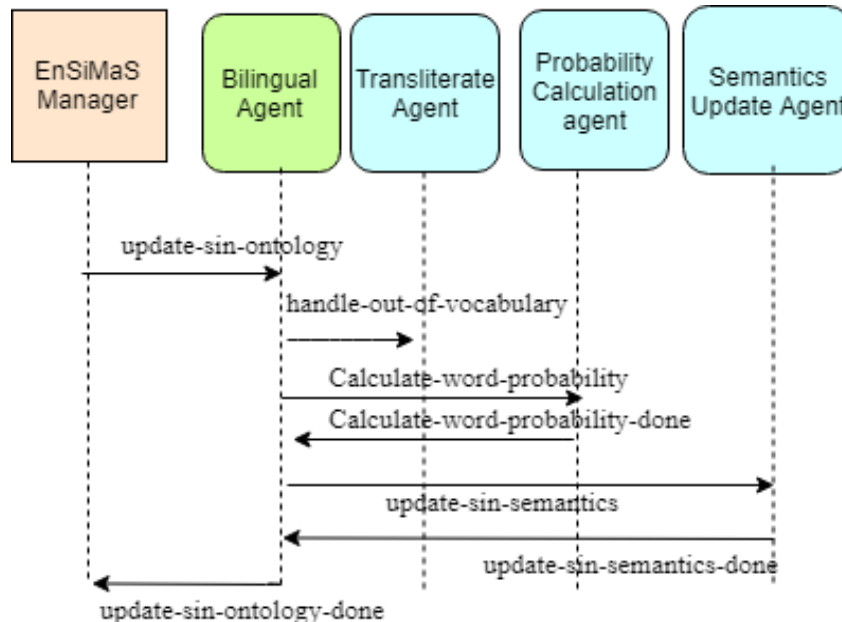


Figure 8.22: Communication diagram of the Bilingual semantics swarm

8.4.5 Sinhala Morphological Generation Swarm

The Sinhala Morphological Generation swarm is used to generate Sinhala Nouns and Verbs according to the given based form and grammar. According to the design of the EnSiMaS, each agent has been designed as a MaSMT agent. Table 8.6 shows the implantation details of each agent in the Sinhala Morphological generation swarm.

Table 8.6: Agents' details of the Sinhala morphological swarm

No	Agent Name	Agent ID	Task
1	SinhalaNounGenerationAgent	generate.501@smg	Generate Sinhala Noun for the given grammar
2	SinhalaVerbGenerationAgent	generate.502@smg	Generate Sinhala Verb for the given grammar
3	SinhalaOtherGenerationAgent	generate.503@smg	Generate Sinhala another word for the given grammar

Sinhala word generation has been done through the 3 agents namely Sinhala Noun Generation Agent, Sinhala Verb Generation Agent and Sinhala Other Generation Agent.

Sinhala Noun Generation Agent: This agent takes noun generation rules from the knowledge base and waiting ready for the word generation. These rules are taken from the BEES project. Figure 8.23 shows the activity diagram of the Sinhala Noun Generation agent.

Sinhala Verb Generation Agent: This agent takes verb generation rules from the knowledge base and waiting for the verb generation. These rules are taken from the BEES project.

Sinhala Other Generation Agent: In general, other words do not participate in the word conjugation. This agent provides relevant Sinhala translation for other words.

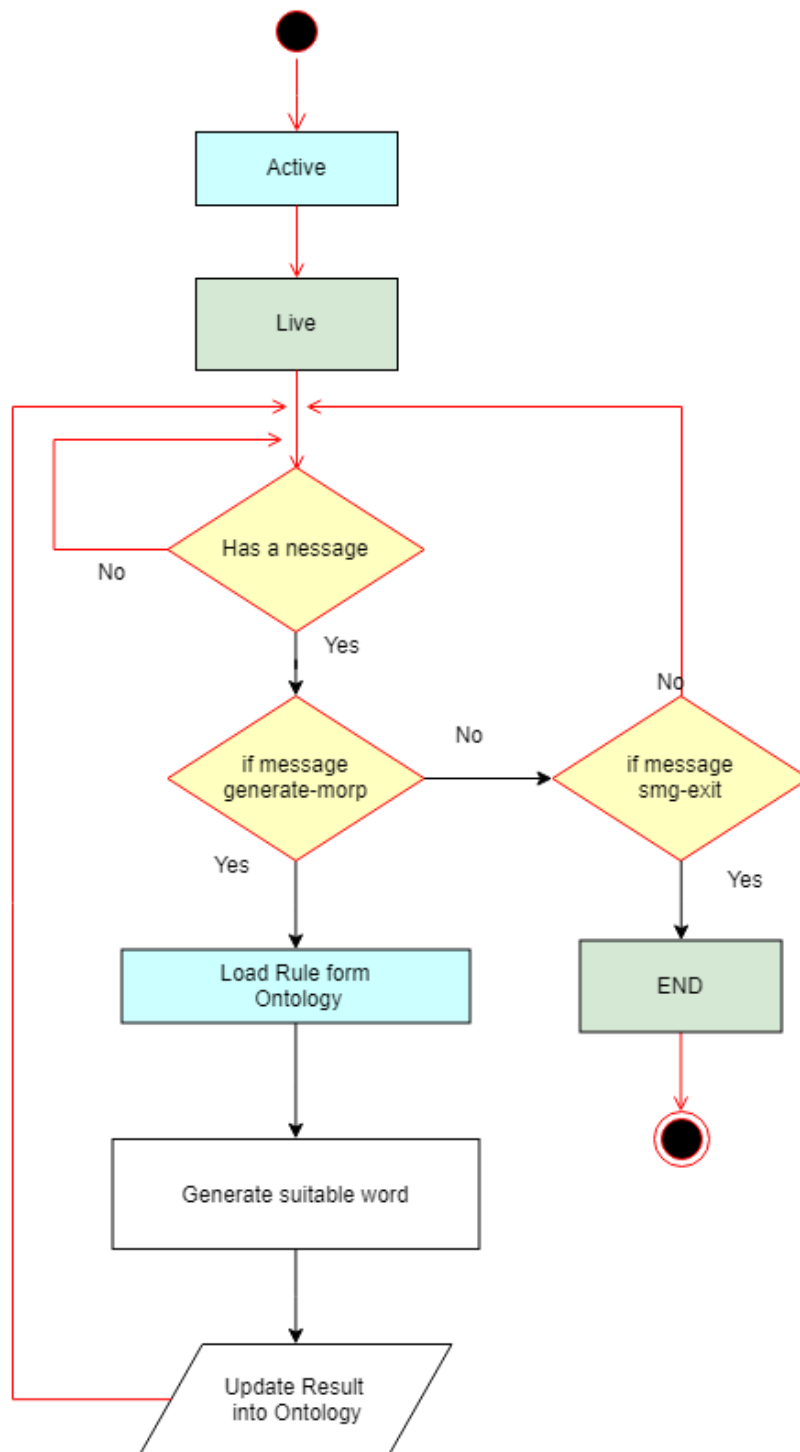


Figure 8.23: Activities of the Sinhala Noun Generation agent

8.4.6 Sinhala Syntactical Swarm

According to the EnSiMAS design, the Sinhala Syntactical Swarm (SSS) should be capable of regenerating suitable Sinhala sentence structure for the existing English sentence structure. The Sinhala syntactical swarm consists of SOVgenerator agent, PPOrderconvertor agent, and Complex Phrase search agent. Table 8.7 shows the agent details of the EnSiMaS Sinhala syntactical swarm.

SOVgeneratoragent: The SOV generator agent is capable of generating suitable grammatically correct Sinhala sentences according to the existing English grammar. This agent takes syntax information from the virtual world.

PPOrderconvertor agent: According to the Sinhala grammar, preposition order takes a different order than the English preposition order. This correction has been done through the PPOrder converter agent.

Complex Phrase search agent: Some phrase in the compound form (one or more phrases) such phrases can be detected through the complex Phrase Search agent.

Table 8.7: Sinhala Syntax Generation Swarm

No	Agent Name	Agent ID	Task
1	SOVgeneratoragent	generate.601@ssg	Generate the SOV order
2	PPOrderconvertor agent	generate.602@ ssg	Generate the preposition phrase word order
3	ComplexPhrase searchagent	generate.603@ ssg	Generate the Complete phrases

8.4.7 Translation Controller Agent

The phrase translation point of view, the translation controller (manager) is the key manager that handles the translation process of the EnSiMaS system. After creating all the required information (after morphological, syntactic, and semantic analysis), the EnSiMaS manager sends a message to the translation manager to start the translation

process. Then the translation manager takes the phrase information from the virtual world and creates Sinhala phrase agents for each English phrase. Note that, according to the type of the English phrase, agents can be categorized into noun phrase agents, verb phrase agents, preposition phrase agents, and other phrase agents. The translation manager should be capable of fully controlling all these agents.

8.4.8 Translation Swarm

The translation swarm is the core swarm of the EnsiMaS system, handles the last and most important stage in the transaction process. Before starting this process, it should be required to complete all the natural language analysis process. Initially, the translation manager takes phrase information from the virtual world and creates the Sinhala phrase agent for each English phrase. These phrase agents are dynamically created according to the English phrase of the input sentence. At the first stage, these Sinhala phrase agents communicate with the virtual world, collect all the required information (English phrase, semantics information etc., which are available on the virtual world), then generate suitable Sinhala phrase(s). For that, it uses the existing English phrase structure and the relevant Sinhala phrase structure for the phrase generation. The phrase structure information can be taken from the virtual world. For the word selection (Sinhala term selection), it uses calculated probability for each word and re-calculates the probability for the generated Sinhala phrase. After the phrase generation top solution (Sinhala phrase that takes the maximum probability), and set of alternative solutions (other Sinhala phrases) are recorded. Then these Sinhala phrase agents communicate with each other agents, (Subject-Verb and Object Verb) phrases and take the selected solution. Figure 8.24 shows the Sinhala phrase agent communication for best phrase pair selection for the subject-verb and object-verb.

Note that this process demonstrates how people search the subject, object, and action verb for the particular sentence and allocate suitable terms according to its context. In here, the suitable context allocation process starts with subject-verb communication.

The subject agent (The agent with a Sinhala subject phrase) communicates with the verb phrase agent (The agent with a Sinhala action verb phrase) and agree in both contexts. After this communication, a suitable Sinhala phrase pair is selected. With communication on verb phrase agent object phrase agent, makes the accepted arrangement for verb and object. Through this communication process, arrangements were not satisfactory (Agents do not agree for the phrase pairs), then the agent should be able to re-arrange their phrases according to the new situation.

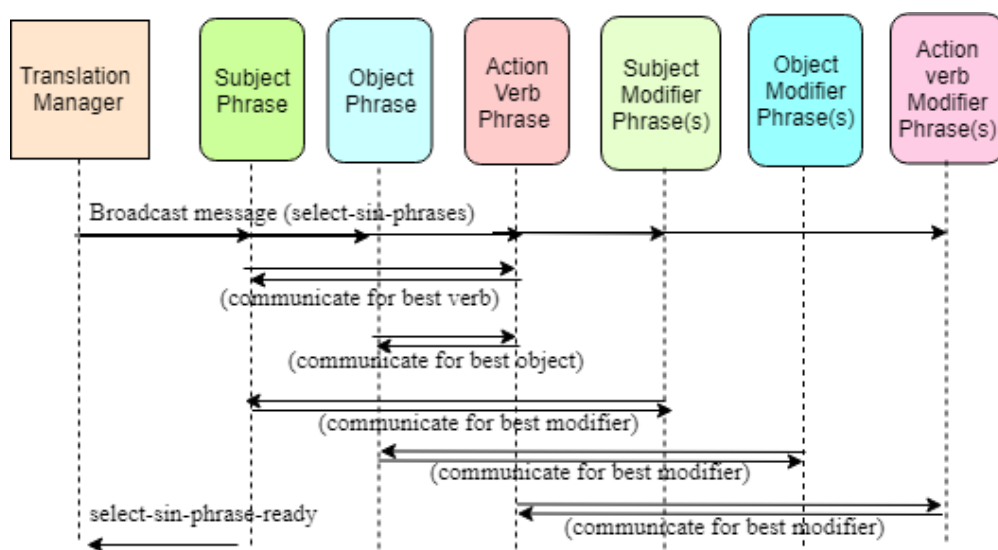


Figure 8.24: Agent communication diagram of the translation swarm

8.4.9 Ontological Swarm

The Ontological swarm has been designed to handle all the information available in the knowledge base and the virtual world. With the support of the ontology managing warm, other agents can access a knowledge base or virtual world easily. This Ontological swarm consists of two types of agents, namely “ontologyAccess” agent and “VirtualWorldAccess” agent. Using Ontology access agent can be used to access relevant information that is available in the knowledge base. The VirtualWorldAccess agent can be used to access the virtual world. Table 8.8 shows the agent details of the ontological swam.

Table 8.8: Ontological Swarm

No	Agent Name	Agent ID	Task
1	OntologyAccessagent	access.701@osm	Communicate with Ontology
2	VirtualWorldAccessAgent	access.702@ osm	Communicate with Virtual World

8.5 Summary

This chapter presented the implantation details of the EnSiMaS. The system has been implemented with two MaSMT controllers and six language processing swarms. Implementation details of each agent have been discussed in this chapter. The next chapter describes how the system has been tested with more details on how the system has been evaluated.

CHAPTER 9

EVALUATION

9.1 Introduction

In the eighth chapter, it was described the implementation details of the EnSiMaS. This chapter presents details of how the EnSiMaS system has been evaluated. For the testing purpose, three applications were developed; namely, EnSiMaS dictionary, EnSiMaS phrase-based sentence editor, and the classical translator. Then the last section of the chapter reports existing MT evaluation methods, the EnSiMaS evaluation process, and the result of the EnSiMaS evaluation.

9.2 English to Sinhala Multi-agent System (EnSiMaS)

This thesis proposed a psycholinguistic approach for machine translation. Hence, the EnSiMaS was implemented as a multi-agent system through the MaSMT framework to simulate the behaviour of human translation. Figure 9.1 shows the top-level GUI of the EnSiMaS system. This GUI gives a way to open the EnSiMaS dictionary, phrase-based editor, or classical translator. More details of the EnSiMaS system tools have been shown under the EnSiMaS user manual (Appendix B).

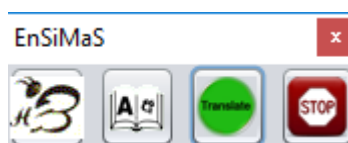


Figure 9.1: Top-level application selection GUI of the EnSiMaS

Using this GUI, users can select any option as they wish. The next few sections briefly describe each component of the EnSiMaS and how they are implemented.

9.3 EnSiMaS Dictionary

To provide an accurate translation, correct lexical resource availability is essential. For that, with the support of various software tools, a sufficient knowledge base has been created. Note that the EnSiMaS bilingual dictionary consists of more than 87000 dictionary words. This dictionary can be used as a stranded dictionary like the Madhura dictionary [262], EnSiTip [173] or Bhasha Dictionary [263]. However, some functionalities have been added, including word correction, grammar editing, and new word insertion. Also, it supports providing a human usage index for each Sinhala term. For instance, an English word has several Sinhala terms; then the system allows the user to select a more suitable Sinhala term from the existing Sinhala terms. The dictionary uses two different types of usage index for each Sinhala term, such as the Google Search index (GSI) and the human usage index (HUI). Through both index values, the system is capable of identifying the best Sinhala translation for the given English term. Further, both GSI and HUI values are used by the EnSiMaS translator to select a more suitable word from the existing Sinhala terms. Figure 9.2 shows the user interface of the EnSiMaS dictionary with the result of the English term “book”.

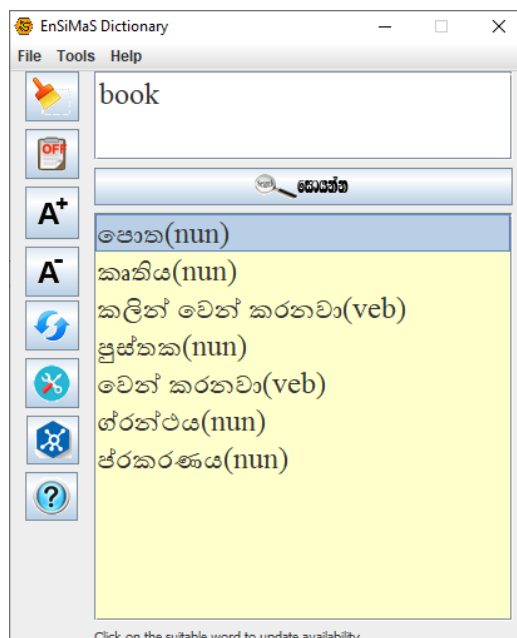


Figure 9.2: GUI of the EnSiMaS dictionary

Note that, this dictionary provides an interface to edit dictionary words (bilingual) and their attributes for better accuracy. The GUI of the EnSiMaS word editor is shown in Figure 9.3.

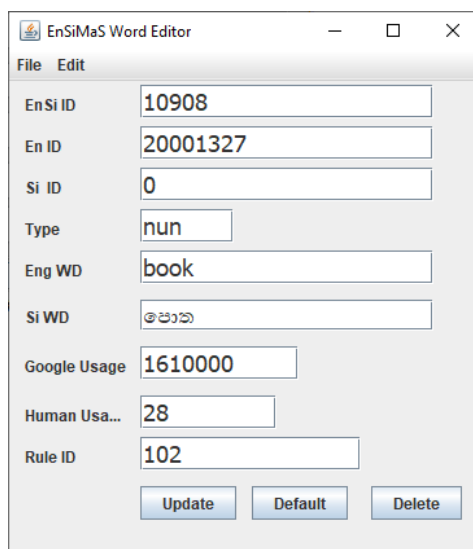


Figure 9.3: A dictionary-based bilingual word editor

9.4 EnSiMaS Translator

EnSiMaS translator is a classical tool for language translation that takes English text and provides grammatically correct Sinhala translation(s). For instance, assume that the system takes “The good boy is reading a new book at the school” as the input sentence. The translation process begins with the EnSiMaS manager taking input text from the GUI and communicating with the ontology manager and creating a new ontology for a translation. Then the EMPS (the agent-based English morphological analyzer) reads each word from the input sentence left to right and analyses each word. The following are the results of the morphological analysis. Here, morphological results are shown with the part of speech TAGs.

```
-1the-- DET
-2good-- RBX JJX JJX NWS
-3boy-- NWS
-4is-- XBF XIS VBP
```


-5reading-- NWS VBG
 -6a-- DET
 -7new-- RBX JJX
 -8book-- VBP NWS
 -9at-- PRP
 -10the-- DET
 -11school-- VBP NWS

After morphological analysis, the SSP (Sinhala semantics processing system) provides available Sinhala words for each English word. For that, SSP uses the existing knowledge base (English-Sinhala bilingual dictionary). If an English word is out-of-the-vocabulary (not in the knowledge base), then transliterate the English word. If this fails, use the same English word. The following information shows the results of semantics processing.

book(8-VBP) කලින් වෙන් කරනවා-veb වෙන් කරනවා-veb
 school(11-VBP) දියුණු කරනවා-veb හික්මවනවා-veb
 book(8-NWS) පොත-nun කෘතිය-nun පුස්තක-nun ග්‍රන්ථය-nun ප්‍රකරණය-nun
 the(1-DET) -det
 good(2-RBX) සාදු-adv
 new(7-RBX) අලුත-adv ලඟදී-adv
 at(9-PRP) දී-prp
 the(10-DET) -det
 good(2-JJX) හොඳ-adj දක්ෂ-adj සොදුරු-adj කලාශා-adj යහපත්-adj හිතවත්-adj ඉටු-adj
 තරමක-adj සුභ-adj ඉෂ්ට-adj
 new(7-JJX) අලුත්-adj නව-adj වෙනස්-adj අමුතු-adj නවීන-adj නුපුරුදු-adj අහිතව-adj දහර-adj
 කෝඩු-adj
 good(2-NWS) යහපත-nun ප්‍රයෝජනය-nun
 boy(3-NWS) පිරිමි ලමයා-nun කොල්ලා-nun ලමයා-nun කුමාරයා-nun කොලුවා-nun බාලයා-nun
 ගැටයා-nun වැඩ කරුවා-nun මානවකයා-nun ආවණේව කාරයා-nun
 reading(5-NWS) කියවීම-nun අර්ථය-nun දැක්වෙන දේ-nun පාඨය-nun පාඨනය-nun වාග්පරිචය-
 nun
 school(11-NWS) විද්‍යාස්ථානය-nun පාසැල-nun විද්‍යාලය-nun පාඨශාලාව-nun විශ්වවිද්‍යාලයේ
 විද්‍යාංශය-nun ශික්ෂාලය-nun
 is(4-VBP) ය-veb
 read(5-VBG) කියවනවා-veb හදාරනවා-veb දක්වනවා-veb ඉගෙනගන්නවා-veb

Then, the syntax analyser (worked as a part of the phrase-based Chanker) enters the process of the translation and identifies each English phrase. The following information shows the results of the syntax analysis.

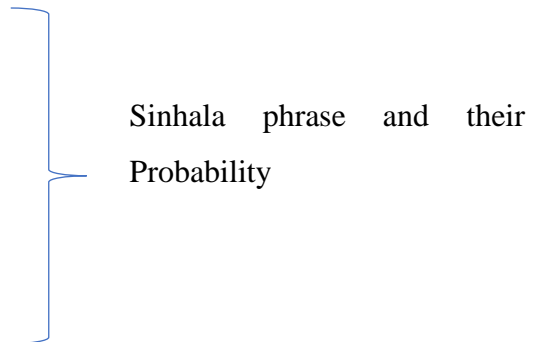
- 1: NP- (1, 3-DET, JJX, NWS) NOS, NWS
- 2: VP- (4, 5-XBF, VBG) ACT-PRC, VBG
- 3: NP- (6, 8-DET, JJX, NWS) NOS, NWS
- 4: PP- (9, 11-PRP, DET, NWS) NOS, NWS

After successful syntax analysis, the translation manager takes all the phrase information and creates Sinhala phrase agents. Note that, Sinhala phrase agents should be capable of generating multiple Sinhala translations for an English phrase, according to the availability of the Sinhala words and considering four factors affecting the language parsing. Note that, the semantics extraction agent on the translation swarm and probability calculation agent in the translation swarm read each translated Sinhala solution and updates the usage of the Google Search index for each Sinhala phrase. The following results describe the language resources available on the Sinhala phrase agent.

1: NP-(1, 3-DET, JJX, NWS) NOS, NWS

හොඳ ළමයා , හොඳ කොල්ලා , හොඳ කුමාරයා , සුභ කොලුවා , හොඳ වැඩ කරුවා , හොඳ බාලයා ,
හොඳ ගැටයා , හොඳ පිරිමි ළමයා , යහපත් මානවකයා , හොඳ ආවතේව කාරයා ,

- 0.5302 : හොඳ ළමයා
- 0.3938 : හොඳ කොල්ලා
- 0.0311 : හොඳ කුමාරයා
- 0.0117 : සුභ කොලුවා
- 0.0098 : හොඳ වැඩ කරුවා
- 0.0095 : හොඳ බාලයා
- 0.0061 : හොඳ ගැටයා
- 0.0058 : හොඳ පිරිමි ළමයා
- 0.0017 : යහපත් මානවකයා
- 0.0004 : හොඳ ආවතේව කාරයා



2: VP-(4, 5-XBF, VBG) ACT-PRC, VBG

දක්වමින් සිටියි , කියවමින් සිටියි , හදාරමින් සිටියි , ඉගෙනගමින් සිටියි ,

0.858 : දක්වමින් සිටියි
 0.0797 : කියවමින් සිටියි
 0.0623 : හදාරමින් සිටියි
 0 : ඉගෙනගමින් සිටියි

4: PP-(9, 11-PRP,DET,NWS) NOS, NWS

විද්‍යාලයේ දී, පාසැලේ දී, පාඨශාලාවේ දී, විද්‍යාස්ථානයේ දී, විශ්වවිද්‍යාලයේ විද්‍යාංශයේ දී, ශික්ෂාලයේදී,

0.8662 : විද්‍යාලයේදී
 0.1281 : පාසැලේ දී
 0.0043 : පාඨශාලාවේ දී
 0.0014 : විද්‍යාස්ථානයේ දී
 0 : විශ්වවිද්‍යාලයේ විද්‍යාංශයේ දී
 0 : ශික්ෂාලයේදී

3: NP-(6, 8-DET,JJX,NWS) NOS, NWS

අලුත් පොතක් , නව කෘතියක් , වෙනස් ග්‍රන්ථයක් , වෙනස් පුස්තකයක් ,
 වෙනස් ප්‍රකරණයක් ,

0.7167: අලුත් පොතක්
 0.1457: නව කෘතියක්
 0.0966: වෙනස් ග්‍රන්ථයක්
 0.0371: වෙනස් පුස්තකයක්
 0.0037: වෙනස් ප්‍රකරණයක්

Considering the Google Search index as a key for calculating the probability and the available context of the generated Sinhala phrase(s), the best Sinhala phrases are identified with the communicating relevant agents in the swarm as required. In here, it takes the phrase pair with the maximum probability of usage. The following list shows the sample probability values for the subject-verb relationship.

#ළමයා කියවනවා	0.579	} Probability of the subject-verb combination
#ළමයා දක්වනවා	0.331	
#ළමයා ඉගෙනගන්නවා	0.047	
#ළමයා හදාරනවා	0.044	
#කොල්ලා කියවනවා	0.727	
#කොල්ලා දක්වනවා	0.175	
#කොල්ලා ඉගෙනගන්නවා	0.07	
#කොල්ලා හදාරනවා	0.029	
#කුමාරයා කියවනවා	0.489	
#කුමාරයා දක්වනවා	0.407	
#කුමාරයා ඉගෙනගන්නවා	0.067	
#කුමාරයා හදාරනවා	0.038	
#කොලුවා කියවනවා	0.683	
#කොලුවා දක්වනවා	0.168	
#කොලුවා ඉගෙනගන්නවා	0.115	
#කොලුවා හදාරනවා	0.034	
#පිරිමි ලමයා කියවනවා	0.694	
#පිරි ලමයා දක්වනවා	0.268	
#පිරිමි ලමයා ඉගෙනගන්නවා	0.031	
#පිරිමි ලමයා හදාරනවා	0.006	

Then the syntax generation swarm takes the information from the virtual world and re-generates the Sinhala sentence by changing the phrase order of the existing phrase list. For multiple solutions, it takes different subjects (Sinhala phrases) and communicates with others with this subject phrase.

Once the system translates the sentence, then thematic information in the sentence is recorded in the virtual world. This information should support extracting pragmatic information from the input text and helping to do a better translation. Figure 9.4 shows the GUI of the EnSiMaS translator.

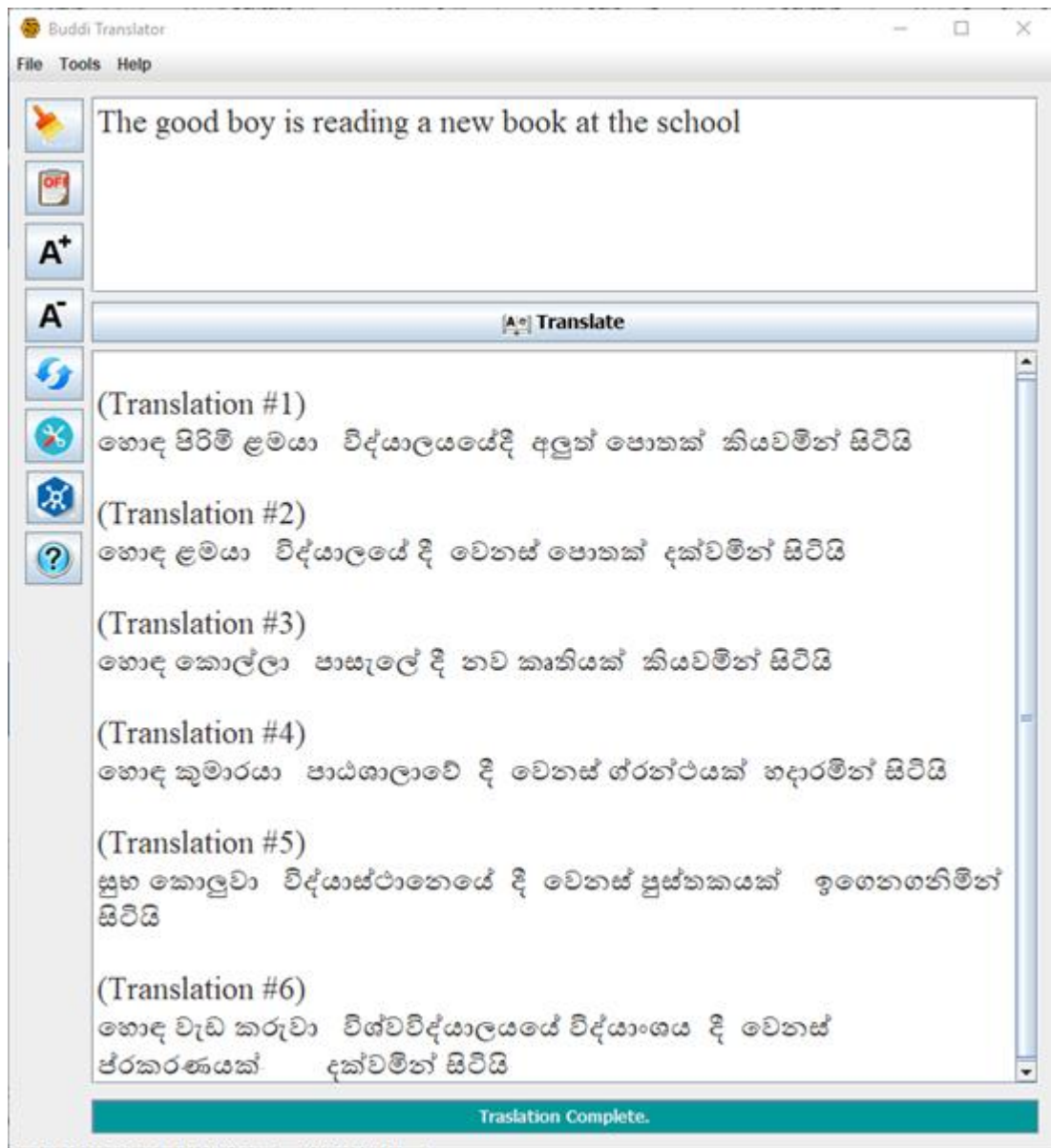


Figure 9.4: GUI of the EnSiMaS Translator

9.5 EnSiMaS Phrase-based Editor

The EnSiMaS phrase-based editor was developed as an intermediate editing tool for language translators (humans) to translate English sentences into Sinhala easily. This editor provides phrase-level editing facilities, including pre- and post-content editing. The main purpose of the EnSiMaS Editor is to capture human knowledge to improve the accuracy of the EnSiMaS. The EnSiMaS editor reads an English sentence and

provides the best Sinhala translation it can provide. Then the editor provides facilitates to changing the appropriate Sinhala phrase from existing phrases. Figure 9.5 shows the GUI of the EnSiMaS phrase-based editor.

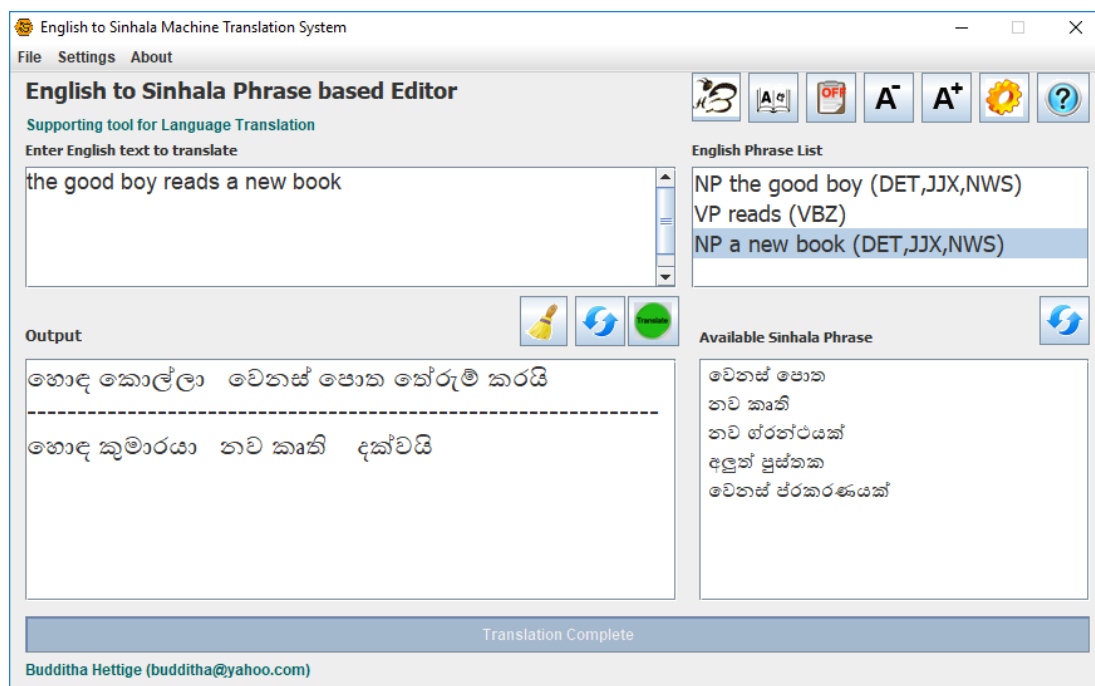


Figure 9.5: GUI of the EnSiMaS phrase-based editor

9.6 Evaluation Strategy of the EnSiMaS

As stated, the hypothesis in this research is that “Multi-agent technology can be used to design a machine translation system capable of processing morphology, syntax and semantics interactively, like humans, rather than sequentially, as current machine translation systems”. In the evaluation of substantiating the above hypothesis, we have used various statistical techniques with their own hypotheses to assess the significance of the conclusion about the hypothesis in the thesis.

Machine translation systems should be capable of translating one language text (source) into the other language (target) without changing its original meaning.

According to the complexity of natural languages, a source language sentence can be translated to several semantically equivalent or similar sentences. Thus, evaluating the quality of machine translation is a complex process that requires a sound knowledge of both languages.

Numbers of evaluation techniques (quantitative and qualitative) are available, including word error rate (WER) [264], sentence error rate (SER) [265], inflexional error rate, etc. All these machine translation system evaluation methods can classify on fully automated evaluations or human-based evaluations [266]. Further, fully automated evaluation can be done through the automatic metrics, with refereeing one or more human reference translations. The section below presents some automated evaluation matrices for machine translation.

9.6.1 Round Trip Translation

The Round-trip Translation (RTT) or back-and-forth translation [267] [268], is one of the earliest methods of evaluating MT systems that translate text from source to target and then translates target to a source using the same machine translation system. The RTT results are closer to the original input sentence; then it says that the quality of the machine translation system is high. However, if the result is not close to the input sentence, then it is impossible to know “where the error occurred” [269]. Thus, some people argue that round-trip translation is not suitable for evaluating MT systems correctly [261].

9.6.2 Word Error Rate

The word error rate (WER) is one of the popular metrics for calculating the accuracy of machine translation systems, especially in speech recognition systems [270].

Through the automated calculation of the word error rate, several applications have been developed for many languages [265]. Note that the WER is derived from the Levenshtein distance (the Levenshtein distance [271] is a string metric for measuring

the difference between two sequences [272]). In general, WER can be computed as the equation given in equation 9.1.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (9.1)$$

where:

S - Number of substitutions in a text,

D - Number of deletions in a text,

I - Number of insertions in a text,

C - Number of correct words in a text,

N – Total number of words in the reference text ($N=S+D+C$)

Note that several algorithms are available to calculate WER automatically and are used to measure the accuracy of the translated results easily [273].

9.6.3 Sentence Error Rate

With comparing the WER, the sentence error rate (SER) is another matrix to evaluate automatic speech recognition systems, as well as output results of the machine translation systems [274]. SER represents the percentage of sentences that are not matched with the reference translation(s).

9.6.4 Translate Error Rate

The translation error rate (TER) [275] is a measure to evaluate an MT system that represents the number of changes (edits) needed to apply a hypothesis to one of the references. The equation 9.2 donates the TER.

$$TER = \frac{\textit{number_of_edits}}{\textit{average_of_reference_words}} \quad (9.2)$$

9.6.5 Inflectional Error Rate

The inflexional error rate (IER) is another error calculation method for the machine translation system that represents “the number of words translated into the correct base form but into the incorrect full form”[276]. Note that the target language is morphologically rich, then the inflexion error may give the wrong results. Therefore, calculating the error of word inflexion helps to make more accurate results of the evaluation. Equation 9.3 shows the IER.

$$IER = \frac{\text{Word With generation Error}}{\text{Total number of inflectional word}} \quad (9.3)$$

9.6.6 BLEU

The BLEU is the most popular inexpensive, fast, language-independent, and automated evaluation matrix for machine translation[277]. In general, the BLEU metric gives "the closer a machine translation is to a professional human translation; the better it is" [278]. The BLEU evaluation matrix provides results in between zero and one, which indicates how similar the candidate text is to the reference text. If the BLEU value is much closer to 1, it represents similar texts.

Consider the process of the BLEU, which calculated scores for individual segments in a sentence. The final score takes considering the average of these scores over the whole corpus. The BLEU score can be calculated using Equation 9.4.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} .$$

Then,

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) . \quad (9.4)$$

However, in general, the BLEU score is used to calculate through the corpus. A modified version is available for the BLEU score with smoothing for the sentence level evaluations [279]. Further, BLUE score calculation programs are freely available on NLTK [280] and GitHub [281].

9.6.7 METEOR

The “Metric for Evaluation of Translation with Explicit Ordering” (METEOR) is another popular evaluation matrix for MT [282]. METEOR has been introduced to overcome existing issues on the BLUE matrix. The “METEOR metric is based on the harmonic mean of unigram precision and recall”. The METEOR MT evaluation matrix gives sentence or segment level good correlation with human judgement. However, the basic unit of both evaluations is the sentence. The METEOR creates an alignment between the candidate translation string and the reference translation string that maps between unigrams. These methods have been applied to many language pairs, including Asian languages like Hindi [283]. Further, there is number of the way on machine translation, therefore selecting the suitable evaluation method is very useful to take the correct accuracy[284].

9.6.8 Human Evaluation

Human evaluation is another way of evaluating MT systems with few disadvantages. Among others, human evaluation is a costly and time-consuming process than automated evaluation [285]. In addition to that, it is difficult to achieve consistency of the same human judge and consistency across multiple judges. Further, the quality of the MT system can calculate in many different ways. Therefore human evaluation standard can be depended on the goal of the evaluation [286]. There are numbers of methods available for human evaluation. Among others, the BEES translator has also been evaluated through human-based methods [287]. Under this BEES evaluation, WER and accuracy have been calculated through human support. However, adequacy-fluency metrics are popular methods for human-based MT evaluation [288].

Adequacy-Fluency Metrics for MT Evaluation

With the above ideas, Koehn and Monz have introduced a common method for human evaluation through adequacy and fluency [289]. Their evaluation method gives way for assigning numerical values to measure quality. This metric is used to evaluate MT performance through adequacy and fluency [290].

Adequacy

The adequacy metric gives “How much of the source information is preserved in the translation?” by considering both source and target. This adequacy judgment asks raters to rate in the scale 1-5 value of meaning expressed in a reference translation (Likert scale [291]). Table 9.1 shows the 1-5 scale of the adequacy metric.

Table 9.1: 1-5 Scale Adequacy matrix

Scale	Adequacy (Meaning translation)
5	All (Perfect)
4	Most
3	Much
2	Little
1	None (No meaning)

Fluency

The fluency matrix gives “How good is the translation regarding the target language quality” by only looking at the target language translation. Table 9.2 shows the 1-5 scale values on the fluency matrix.

Table 9.2: Fluency value in the Likert scale

Scale	Fluency (According to the target translation)
5	Flawless translation
4	Good translation
3	Non-native translation
2	Diffluent translation
1	Incomprehensible translation

With these methods, different ranks are obtained. After raters rate the sentence, it should be required to check the agreement in between each rater and finally measure the ordinal association between measured quantities through the Kappa coefficient or Kendall's rank coefficient [292].

Agreement between different raters

After rating each sentence, it should require to take annotator agreement for adequacy and fluency tasks, which can be measured using the “Kappa coefficient” [292]. The “Kappa coefficient” is a statistical measure that is used to measure “inter-rater” reliability for qualitative items [293] [294]. Note that more research gives attention to the adequacy than the fluency [295]. Equation 9.5 gives the Kappa coefficient.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (9.5)$$

Where: P(A) is the proportion of times annotators agree, and P(E) is the proportion of times annotators are expected to agree by chance” (5-point scale → p(E) = 1/5).

The Kappa coefficient can be used to agree on what constitutes good or poor levels. Table 9.3 shows the Kappa values for agreement.

Table 9.3: Fleiss' Kappa values for agreements

Kappa (K)	Agreement
K < 0	No agreement
0.00 – 0.20	Poor agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Good agreement
0.81 – 1.00	Very good agreement

However, Kappa coefficient values are taken only for two raters. For more raters, Fleiss' kappa can be used. "Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items". Fleiss' kappa can be taken from Equation 9.6.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (9.6)$$

Where:
$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1)$$
 and
$$\bar{P}_e = \sum_{j=1}^k p_j^2$$

In general, Fleiss' kappa will be higher when the agreement is perfect. Table 9.4 shows Fleiss' kappa values for agreements.

Table 9.4: Fleiss' Kappa values for agreements

Fleiss' kappa (K)	Agreement
K < 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Measure the ordinal association between reference translation and a system-generated translation

It is better to measure the ordinal association between the reference translation and a system-generated translation. This evaluation needs to do with the measure of the adequacy and fluency matrix. For that, Kendall's tau correlation coefficient can be

used. Kendall's tau correlation coefficient is a statistical measure used to measure the ordinal association between two measured quantities.

The expression for “Kendall's rank coefficient” is given in Equation 9.7:

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j) \quad (9.7)$$

Note that, “Kendall's rank coefficient” [296] ranking value ranges from -1 to +1. If there is good agreement between the two rankings, then the value is 1 and disagreement between the two rankings then the value is -1. Further, if two rankings are independent, then the ranking value would be approximately zero.

The above-mentioned statistics theories and methods have been used in the EnSiMaS evaluation. The rest of the chapter describes the experimental setup, data analysis, and results of the evaluation clearly.

9.7 Experiment

EnSiMaS was evaluated with the following procedure:

1. As the first step, 85 sample sentences were created by considering the different structure, including different types of tenses and commonly used sentence patterns (different types of sentences).
2. Each sentence was given to the subject expert (professional language translators) to translate into Sinhala. This translation has been considered as the reference translation.
3. Through the EnSiMaS System, each sentence has been translated into Sinhala. Note that, the best three translated sentences were taken for each input English sentence. For the evaluation purpose, 85 English sentences, 255 Sinhala translations, and 85 reference translations were used.

4. With these records, the WER, SER, and IER were calculated.
5. Using the existing Python program, the BLEU score was calculated for all solutions separately considering the reference translations.
6. The evaluation form has been created for the randomly selected 25 sample sentences. With this evaluation form, adequacy and fluency rates are taken from 55 raters.
7. Through the statistical analysis, raters' agreements and correlations were taken.

Table 9.5 shows the selected 25 sample sentences and some translations for each sentence.

Table 9.5: Sample 25 sentences with translated results

No	English source	EnSiMaS Translation
1	I am writing a book	මම පොතක් රචනා කරමින් සිටිමි
2	The good boy reads a new book	හොඳ කොල්ලා අලුත් පොතක් කියවයි
3	A good student will eat rice at the canteen	හොඳ ශිෂ්‍යයෙක් ආපන ශාලාවේ දී බත් අනුභව කරන්නේය
4	The man with his wife went to the party	පුරුෂයා සහ ඔහුගේ සහකාරිය උත්සවයට ගියේය
5	The good boy was eating an apple at the school	හොඳ ළමයා විද්‍යාලයේ දී ඇපල් ගෙඩියක් අනුභව කරමින් සිටියේය
6	The good boy and a beautiful girl have written an essay	හොඳ කුමාරයා ද ලස්සන කුමරියක් වාක්‍ය රචනාවක් රචනා කරලා තියෙයි
7	I will write a story for my children	මම මගේ ළමයි සඳහා කතාවක් රචනා කරන්නෙමි
8	I play basketball every week with my friends	මම මගේ මිතුරු සමඟ සෑම සතියකම පැසිපන්දු ක්‍රීඩා කරමි
9	My good friend was singing a song at the school	මගේ හොඳ මිතුරා විද්‍යාලයේ දී ගීතයක් ගායනා කරමින් සිටියේය
10	A boy and a girl sing a song on the bus	පිරිමි ළමයෙක් සහ ගැහැණු ළමයෙක් බස් රථයේ ගීතයක් ගායනා කරති

11	The boy is going to eat rice with my friend	කොල්ලා මගේ මිතුරා සමග බත් අනුභව කරන්නේය
12	I wrote a letter at the school	මම පාසලේදී ලිපියක් රචනා කළෙමි
13	We read newspapers daily	අපි ප්‍රවෘත්ති පත්‍ර දිනපතා කියවමු
14	She can ride a bicycle and drive a car	ඇයට බයිසිකලයක් පැදගෙන හා වාහනයක් පදවන්න හැකිය
15	Mother prepares the breakfast at the kitchen for her children	මව කුස්සියේ දී ඇයගේ ළමයි සඳහා උදෑසන ආහාරය සුදානම් කරයි
16	My dog ate rice with meat	මගේ බල්ලා මාංස සමග බත් කෑවේය
17	I was eating a big pizza with my friends	මම මගේ මිතුරන් සමග වැදගත් ඉතාලියානු ආහාර විශේෂයක් කමින් සිටියෙමි
18	I am selling my motorcycle and buying a new car	මම මගේ බයිසිකල විකුණනවා සහ නව රථයක් මිලට ගන්නවා
19	We wrote a book	අපි පොතක් රචනා කළෙමු
20	My dog and his cat are eating rice with meat	මගේ බල්ලා සහ ඔහුගේ පුසා මාංශ සහ බත් කමින් සිටියි
21	A clever student reads good newspapers daily	දක්ෂ ශිෂ්‍යයෙක් හොඳ පුවත්පත් දිනපතා කියවයි
22	The singer has sung a new song	ගායකයා නව සින්දුවක් ගායනා කරලා තියෙයි
23	The neighbour bought a radio	අසල්වැසියා රේඩියෝවක් මිලට ගත්තේය
24	We will draw a painting at every weekend with our friends	අපි සෑම සති අන්තයකදීම සමග අපගේ මිතුරන් සමග සිත්තමක් අඳිමු
25	The strongest rain ever recorded in India shut down the financial hub of Mumbai, snapped communication lines, closed airports and forced thousands of people to sleep in their offices or walk home during the night, officials said today.	India හිදී ලියා තැන්පත් කරන ලද දැඩි වර්ෂාව Mumbai හි මුදල් පිළිබඳ මධ්‍යස්ථානය වහසි සන්නිවේදන මාර්ග ආවරණය කළේය ගුවන්තොටුපළ සහ ඔවුන්ගේ කාර්යාලය හිදී ජනතාව නින්දට හෝ රාත්‍රිය අතරතුර ගමන් කර නිලධාරියෝ අද ප්‍රකාශ කළේය

9.8 EnSiMaS vs Google Translator

This section gives a brief comparison of Google translator and EnSiMaS by considering their accuracy and features. Google provides free language translation service among more than 100 languages, named “Google Translator”. In general, the quality of the Google translation is depended on amounts of data and, based on previous translations [297]. Many researchers have noted that this tool may be the perfect travel companion, it’s not a reliable language service[298]. As such, there are many reasons not to rely on Google Translator [299] [300] [301]. Below are some commonly cited reasons.

1. Google recommends not to rely completely on its free translation service
2. Translation causes hilarious situations
3. Google Translator doesn’t have a proofreading service
4. Google Translator has limitations in grammatical correctness in some translations
5. Translation quality is depended on amounts of data

In comparison with Google Translator, EnSiMaS translation provides grammatically correct translation, and it provides more than one translation depending on the availability of lexical resources. Further EnSiMaS provide correction facilities with its Phrase-based editor. Table 9.6 shows Features comparison between EnSiMaS vs Google

Table 9.6: Comparison between EnSiMaS vs Google Translator

System	Approach	Grammatically Correct Translation	Multiple Solutions	Correction (Human Editing)	Update Facilities	Speed	Bilingual Translation
Google Translator	Statistical/ Neural Machine Translation	X	X	X	X	√	√
EnSiMaS	Hybrid (MAS + psycholinguistic)	√	√	√	√	X	X

9.9 Results and Data Analysis

The EnSiMaS system was analysed using 85 English sentences. Each English sentence was translated into Sinhala using EnSiMaS, and 255 Sinhala translations were taken. In addition to that, reference translations were taken with a human expert's support.

9.9.1 Details of the sample set

85 sentences are used to evaluate the system. These sentences are taken from the internet and a book on English linguistics for language teaching [302]. The descriptive statistics of the input sentences are given in Table 9.7.

According to the sample used minimum word count for the sample, is 4 and the maximum word count of the sample sentence is 38. The mean value of the sample sentence set is 9.68. it refers that most sentences have more than 9 words.

Table 9.7 Summary of descriptive statistics of the 85 input sentences

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Number_of_words	85	4	38	9.68	4.953
Valid N (listwise)	85				

Also, the distribution of the number of words among each input English sentence is shown in Figure 9.6.

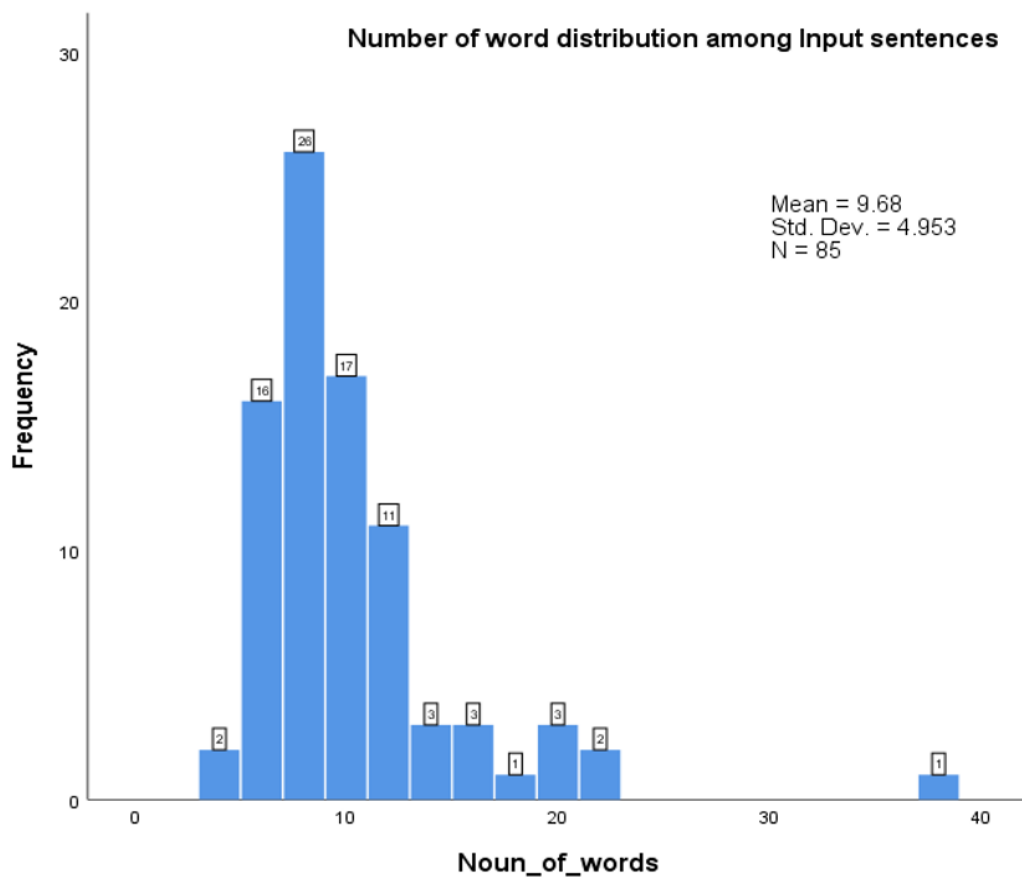


Figure 9.6: Distribution of the number of words among input sentences

Using the translated output (three translations for each sentence), WER, IER and SER were calculated. Table 9.8 shows the results of the word error rate, inflexion error rate and sentence error rate.

Table 9.8 Calculated WER, IER and SER for the translations

EnSiMaS Translation	WER	IER	SER
Solution 1	5.26%	5.26%	4.37%
Solution 2	10.53%	7.02%	5.21%
Solution 3	10.53%	7.89%	6.54%

Further, with reference translation, the BLEU score was calculated for the three translations. Table 9.9 shows the calculated BLEU score for each solution.

Table 9.9 Calculated BLEU results for each translation

EnSiMaS	P(1)	P(2)	P(3)	P(4)	BLEU
Solution 1	0.92897727	0.89141004	0.87358490	0.87358916	0.89160756
Solution 2	0.70738636	0.55915721	0.46037735	0.40180586	0.52009204
Solution 3	0.60653409	0.46029173	0.38679245	0.33408577	0.43581893

9.9.2 Adequacy and Fluency

Adequacy and fluency of the best solution (Translation 1) were evaluated with human support. The evaluation form was used to evaluate a randomly selected sample of 25 sentences from the existing 85 English sentences. The evaluation form was given to raters, and they were asked to rate according to the translated sentences' adequacy and fluency. Using the evaluation results, Fleiss' kappa coefficient was calculated for each adequacy and fluency agreement. Table 9.10 shows the summary for Fleiss' kappa coefficient values for adequacy and Table 9.11 shows the summary for Fleiss' kappa coefficient value for fluency.

Note that:

Hypotheses for the raters' agreements on adequacy can be stated as

H₀: There is no agreement on adequacy between each rater.

H₁: There is an agreement on adequacy between each rater.

The agreement was calculated through the Fleiss' kappa coefficient. Calculated results show that the overall Fleiss' kappa coefficient is 0.277 with 0.000 significance. (Thus, H₀ can be rejected at 5% level of significance.)

Therefore, it can be concluded that there is a fair agreement between raters on their adequacy ratings.

Table 9.10 Fleiss' kappa coefficient values for Adequacy

Overall Kappa						
	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
Overall	.277	.009	30.824	.000	.259	.295

Kappa's for Individual Categories							
Rating Category	Conditional Probability	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
3	.458	.386	.012	33.393	.000	.363	.408
4	.411	.121	.012	10.516	.000	.099	.144
5	.718	.370	.012	32.065	.000	.348	.393

Hypotheses for the raters' agreements on fluency can be stated as

H₀: There is no agreement on fluency between rater

H₁: There is an agreement on fluency between rater

The agreement was calculated through the Fleiss Kappa coefficient. Calculated results show that the overall Fleiss' kappa coefficient is 0.308 with 0.000 significance. (Thus H₀ can be rejected at 5% level of significance.)

Therefore, it can be concluded that there is a fair agreement between raters.

Table 9.11 Fleiss' kappa coefficient values for Fluency

Overall Fleiss' kappa coefficient values for Fluency						
	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
Overall	.308	.008	38.563	.000	.292	.323

Kappa's for Individual Categories							
Rating	Conditional Probability	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
2	.608	.589	.012	50.973	.000	.566	.611
3	.277	.178	.012	15.418	.000	.155	.201
4	.454	.293	.012	25.375	.000	.270	.316
5	.732	.322	.012	27.905	.000	.300	.345

Considering all the raters' results, the percentage value for the level of adequacy and fluency was calculated for translation #1. Figure 9.7 represents the percentage distribution of both adequacy and fluency.

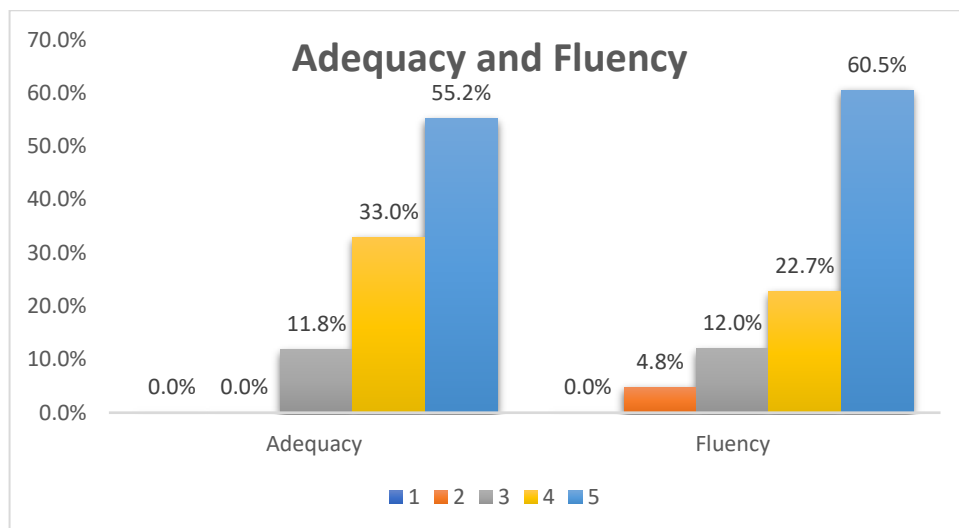


Figure 9.7: Percentage distribution on adequacy and fluency values for EnSiMaS Best Translation

Besides, adequacy levels and fluency levels for five randomly selected raters are shown in Figure 9.8 and Figure 9.9. According to the figures, it can be concluded that more than 80% of raters (higher proportion) has given for the higher level (4 and 5) for adequacy and more than 60% rates have given for the higher value of fluency.

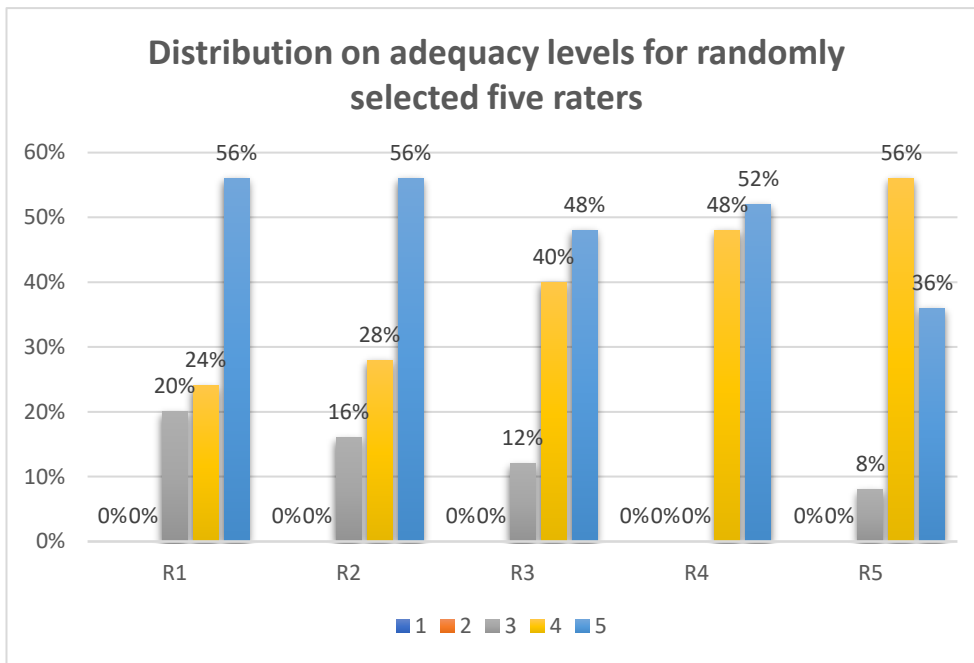


Figure 9.8: Distribution on adequacy rates on five different raters

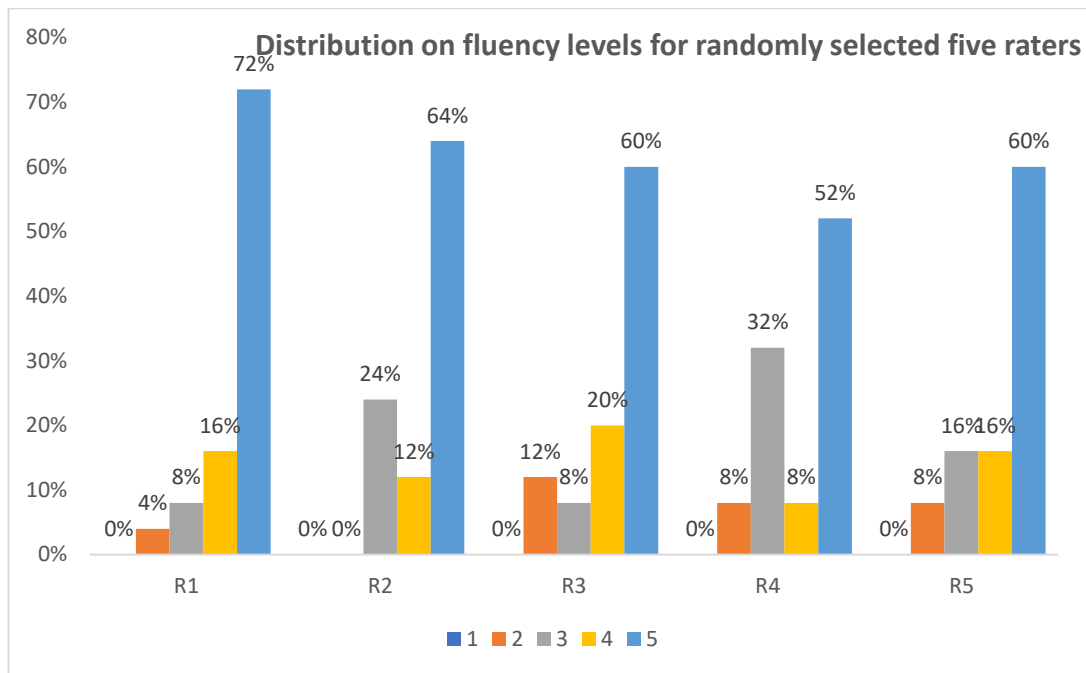


Figure 9.9: Distribution on fluency rates on five different raters

Measure the ordinal association between human translated results and EnSiMaS translation

Ordinal association between human translation and EnSiMaS translation on adequacy is given below.

Hypotheses for the association between human translation (reference translation) and EnSiMaS translation on adequacy can be stated as follows:

H_0 : There is no association between the adequacy of human translation and adequacy of EnSiMaS translation.

H_1 : There is an association between the adequacy of human translation and adequacy of EnSiMaS translation.

Association has been calculated through Kendall's rank correlation coefficient. Calculated results show that Kendall's rank correlation coefficient is 0.154 with 0.005 significance. (Thus, H_0 can be rejected at 5% level of significance). Therefore, it can be concluded that there is a fair association between adequacy of human translation

and adequacy of EnSiMaS translation. Table 9.12 shows the test results of “Kendall's rank correlation coefficient.”

Table 9.12: Summary of the Kendall's rank correlation coefficient for adequacy between Human translation and EnSiMaS translation

Correlations on Adequacy			
		Adequacy of the Reference Translation	Adequacy of the Machine Translation
Adequacy of the Reference Translation	Pearson Correlation	1	.154**
	Sig. (2-tailed)		.004
	N	350	350
Adequacy of the Machine Translation	Pearson Correlation	.154**	1
	Sig. (2-tailed)	.004	
	N	350	350

Ordinal association between human translation and EnSiMaS translation on fluency is given below.

Hypotheses for the association between human translation (reference translation) and EnSiMaS translation on fluency can be stated as follows:

H₀: There is no association between the fluency of human translation and fluency of EnSiMaS translation.

H₁: There is an association between the fluency of human translation and fluency of EnSiMaS translation.

Association has been calculated through Kendall's rank correlation coefficient. Calculated results show that Kendall's rank correlation coefficient is 0.483 with 0.000 significance. (Thus, H₀ can be rejected at 5% level of significance) Therefore, it can be concluded that there is a moderate positive association between fluency of human

translation and fluency of EnSiMaS translation. Table 9.13 shows the test results of “Kendall's rank correlation coefficient”.

Table 9.13: Summary of the Kendall's rank correlation coefficient for fluency between Human translation and Fluency on EnSiMaS Translation

Correlations on fluency matrix			
		Fluency of the Reference Translation	Fluency of the Machine Translation
Fluency of the Reference Translation	Pearson Correlation	1	.483**
	Sig. (2-tailed)		.000
	N	350	350
Fluency of the Machine Translation	Pearson Correlation	.483**	1
	Sig. (2-tailed)	.000	
	N	350	350
**. Correlation is significant at the 0.01 level (2-tailed).			

9.10 Conclusion of the Data Analysis

According to the Fleiss' Kappa coefficient, the calculated value for adequacy levels (0.277) reveals that there is a significantly fair agreement on adequacy between raters. Also, the Fleiss' kappa coefficient calculated for fluency levels (0.308) reveals that there is a significantly fair agreement on fluency between raters. Also, Kendall's rank correlation coefficient (0.154) shows that there is a weak positive association between adequacy levels of human translations and system translations. Calculated Kendall's rank correlation coefficient (0.483) for fluency levels of human translation and system translation reveals that there is a moderately positive association between fluency levels of human translation and the EnSiMaS system translation.

9.11 Summary

The first part of the chapter reports testing tools that have been used to test the EnSiMaS system, namely EnSiMaS Dictionary (ontology editor), the phrase-based editor, and the classical translator. Then the next section of the chapter reports the evaluation process of the EnSiMaS. The EnSiMaS has been tested with 85 sample sentences. With these results, the WER, IER, SER and BLEU scores were calculated. In addition to that, adequacy and fluency were ranked from 55 raters. Results were statistically analyzed.

CHAPTER 10

CONCLUSION AND FURTHER WORK

10.1 Introduction

This thesis proposed a novel approach to machine translation, which is based on the concepts behind human language translation. The previous chapter reported the evaluation methods, including evaluation results of the EnSiMaS. This chapter reports a briefing of the thesis, including revisit note on the proposed approach, conclusion, objective-wise achievements, limitations and future works of the research.

10.2 Hybrid Approach for Machine Translation

Machine translation systems are now commonly accepted by a large community of people for translating one language text into another as a quick and low-cost solution. However, the existing system has some quality gaps in-between machine-generated output and the human-translated well perfect translations. To reduce the quality gap between human translated results and machine-translated solutions is the current trending research direction.

Note that, at the early days of the machine translation systems were introduced, concepts on the human translation were investigated. However, at that time, computing technologies are not sufficient to simulate such human translation process (Summary was taken from the first MT conference in 1952). With the power of the Multi-agent technology, this research proposed a new machine translation approach, that should capable to translate like a human (Use psycholinguistics concepts for machine translation) This can be considered as the primary contribution on this research for the field of computing especially on Machine Translation.

The proposed approach is based on psychological language parsing techniques used by humans to understand a text specially written in the English language [31]. Theoretically, this proposed MT approach can be demonstrated with the model on the garden path model and the constrained stratification model. This proposed approach

takes concepts on both theories and makes new concepts, which can be implemented through the multi-agent system technology. According to the approach, it collects all the possible solutions (like constrained stratification) and assigns context through subject-verb, verb-object agreement (like the garden path model). Further, translation has been done through the phrase-based, then subject-verb agreement and object-verb agreement were taken, all have done through the agents' communication. The proposed translation approach has been tested through the multi-agent system, EnSiMaS. For the testing purpose, three systems were developed. The EnSiMaS dictionary is a bilingual dictionary tool, which was used to customise the knowledgebase for English and Sinhala languages. The phrase-based editor is another tool that was used to enhance the accuracy of the phrase translation with human support. The classical translator can be used to translate one or more sentences (sentence or phrase).

The EnSiMaS system was tested with 85 sample English sentences. For each English sentence, three different translations were taken. According to the evaluation result, WER, IER and SER were calculated. Also, the BLEU scores were calculated for each translation. Finally, adequacy and fluency rates were taken from 55 human evaluators considering the human-translated reference sentences. According to the statistical analysis, there is a weak positive association between adequacy levels of human translations vs EnSiMaS system translations and a moderate positive association between fluency levels of human translation and EnSiMaS system translation.

10.3 Conclusion

The hypothesis of the research is stated as multi-agent technology can be used to design a machine translation system, capable of processing morphology, syntax and semantics interactively, like humans. Therefore, a novel machine translation approach was proposed by combining the existing two human language parsing concepts, namely the garden path model and the constraint satisfaction model. The approach has been simulated through the multi-agent system technology.

Note that most of the existing approaches provide unidirectional interaction among morphological, syntactical, and semantic processing. The proposed hybrid approach avoids these unidirectional language processing concepts and proposed interactive communication and made an agreement through the agents. In addition to that, according to the power of the multi-agent systems, most of the tasks can be done parallel when they can do parallel. However, to increase the performance of the system, some analysis is required to complete (morphological and syntactical analysis have been done before going to translation). Then agents can interactively communicate with each other and solutions were taken.

This hybrid approach has been simulated through the Multi-agent system EnSiMaS, capable to translate English sentence into Sinhala. EnSiMaS translator has been tested with 3 subsystems namely EnSiMaS dictionary, EnSiMaS phrase-based editor and EnSiMaS translator.

The performance of the EnSiMaS system was tested using 85 English sentences and 255 Sinhala translations. According to evaluation, 5.26% Word Error Rate, 5.26% Inflection Error rate, 4.37% sentence error rate and 0.89 Bleu scores were obtained. The adequacy and fluency of the best solution were evaluated through human support. 88.2% raters rates for most or perfect translation on adequacy and 83.2% raters rates for good or flawless translation on fluency. According to the statistical analysis, there is a fair agreement between raters on their adequacy and fluency ratings. Besides, the ordinal association between human translation and EnSiMaS translation on adequacy and fluency were taken. A fair association between adequacy of human translation and adequacy of EnSiMaS translation were obtained. Further, a moderate positive association between fluency of human translation and fluency of EnSiMaS translation were taken.

Therefore, it can conclude that the proposed interactive approach can work to provide language translations like humans successfully and Multi-agent system technology can also be used to machine translation successfully.

10.4 Objectives-wise Achievement

This research aimed to implement the human translation approach into fully automated machine translation through the multi-agent system technology. The following objectives mentioned in the first chapter were completed to achieve this goal.

Objective 1: Critically review of existing machine translation approaches and systems

The second chapter critically explains the state of the art on machine translation, including an in-depth study on concepts behind machine translation, its historical development, popular approaches, techniques and selected systems on machine translation, covering more than 300 references. Among existing MT approaches, the rule-based approach has been identified as an early approach to MT and is still used for many systems, especially for low resources languages. At present, statistical and machine learning approaches (Neuro-linguistics approach) have been considered as the most popular and successful approaches for machine translation. However, those approaches required much language resources to take success. Therefore, low resource languages like Sinhala take a bit of difficulty to use said approaches because of the low resources. Further, chapter four also briefly described some related natural language processing techniques that are used in the English to Sinhala Machine Translation, including related techniques on morphological processing, syntax processing, and semantic processing.

Objective 2: An in-depth study to model Sinhala and English languages to build an ontology for agents

Agents need language-specific knowledge for the agents' ontology. Thus the ontological model has been designed for the Sinhala and English languages that can be directly used to model the MT requirements. According to this ontological model, a word has been identified as a basic unit of language. In addition, the fundamental

unit of the meaning has been identified as a phrase. The sentence has been structured with numbers of phrases that preset the required meaning. With this idea, the object-oriented based ontological model has been designed and presented in chapter eight, considering the in-depth study of both Sinhala and English languages. Further, chapter three also reported morphology, syntax and semantics on English and Sinhala languages.

Objective 3: Critically review MAS technology for MT

Chapter seven reported the required details on multi-agent system technology, including fundamental concepts on MAS, with the concepts of different type of the agents and communication methods also reported in the seventh chapter. In addition to that, selected MAS development frameworks were also presented, including their features and limitations. Compared with existing frameworks, JADE provides a fully distributed environment for MAS development. However, JADE takes more resources, and it is challenging to implement a large number of dynamically created agents that required to agent-based machine translation. Thus, a new framework (MaSMT) was designed and developed considering such requirements.

Objective 4: Define a Language Translation method that is capable of translating like a human

The sixth chapter reported the proposed approach. This approach is based on the concept “How human translate a sentence in a better way than the machine.” Therefore, psycholinguistic language translation techniques, namely the “garden path model” and the “constraint satisfaction model”, were considered to model this proposed approach for machine translation. The proposed translation model was implemented through the EnSiMaS, which was described in the eighth and ninth chapters.

Objective 5: Design and develop EnSiMaS using multi-agent system technology

The sixth chapter discussed the approach proposed for the machine translation, including the theory behind the approach, input, output, and process of the translation. Then chapter seven and eight explained the design and implementation of the EnSiMaS system. To achieve performance for the agent-based machine translation, the MaSMT framework was developed. The MaSMT has been developed using the modified AGR organizational model that provides an infrastructure of the agents, communication methods for agents, agent status controlling, and a tool for agent monitoring. Through the MaSMT framework, EnSiMaS was implemented to translate English text into Sinhala. Then Chapter 9 also gives a brief demonstration of the translation procedure of the EnSiMaS using the proposed approach.

Objective 6: Evaluate the system

Chapter nine discussed the experimental setup of the system with steps following for the evaluation. For the testing purpose, three applications were developed, namely, EnSiMaS dictionary, EnSiMaS phrase-based sentence editor, and the classical translator. Through the EnSiMaS translator, an evaluation was done with human support. As the first steps, the EnSiMaS system was tested using 85 English sentences. Each English sentence was translated using EnSiMaS, and 255 Sinhala translations were taken. Using the translated results (three translations for each sentence), WER, IER, SER and BLEU scores were calculated. Then the adequacy and fluency of the best solution (in here translation #1) were evaluated through human support. The evaluation form was used to evaluate with randomly selected 25 samples from the existing 85 English sentences. With the experimental results, Kendal's tau correlation coefficient was used to check the correlation between human translations and system translations. Fleiss' kappa coefficient of each adequacy and fluency agreements were also calculated.

10.5 Limitations

This thesis proposed a psycholinguistic-based hybrid approach for MT. The proposed approach was tested through the multi-agent system named EnSiMaS. This system has several limitations. The present translation system can only work for unidirectional translation (English to Sinhala). However, it can be by-directional by improving the language processing modules on both languages. Besides, the present EnSiMaS tool does not support idiomatic phrase translations. Including the idiomatic phrase as a block, it can be achieved.

10.6 Further Work

As a further work of the research, the EnSiMaS system can be improved as a bidirectional translation for English-Sinhala by including the required language processing modules for the Sinhala and English languages such as English morphological generator English composer (Syntax generator) Sinhala to English phrase-based phrase translator, Sinhala morphological analyzer and Sinhala parser. Sinhala is a morphologically rich language than English. Therefore Sinhala language analysis bit difficult than the English language analysis.

EnSiMaS can be enhanced to provide a solution to “idiomatic translation” by adding such resources into the knowledge base. Translation of the idiomatic phrases are challengeable and it required to take the hidden meaning than what they appear.

Introducing concepts dictionary and handle phrase-level semantics is another research direction of the project (At present EnSiMaS provides multiple translations). Further, the EnSiMaS offers a grammatically correct set of translations for each phrase. Those translated phrases (Sinhala translation for particular English Phrase) can be taken as the language resources (parallel corpus for English-Sinhala) for the NMT or statistical-based machine translation, with human editing as required.

10.7 Summary

This last chapter reported the conclusions and further works of the research, including each objective and limitations of the research. Further, this chapter also points out some important new research directions.

References

- [1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Pearson education, 2005.
- [2] “The History of Language Translation”, 08-May-2018. [Online]. Available: <https://www.unitedtranslations.com/great-history-of-language-translation/>. [Accessed: 03-Sep-2019]
- [3] “Kumarajiva | Buddhist scholar,” *Encyclopedia Britannica*. [Online]. Available: <https://www.britannica.com/biography/Kumarajiva>. [Accessed: 23-Dec-2018]
- [4] “History of Bible Translations” [Online]. Available: <http://www.historyworld.net/wrldhis/PlainTextHistories.asp?historyid=ac66>. [Accessed: 14-May-2019]
- [5] *Language and Machines: Computers in Translation and Linguistics*. Washington, D.C.: National Academies Press, 1966 [Online]. Available: <http://www.nap.edu/catalog/9547>. [Accessed: 27-Nov-2019]
- [6] “From The Business Of Language To The Language Of Business: The Future Of Translation Worldwide.” [Online]. Available: <https://www.digitalistmag.com/future-of-work/2018/05/17/future-of-translation-worldwide-06168565>. [Accessed: 23-Dec-2018]
- [7] “How many languages in the world are unwritten?,” *Ethnologue*, 09-May-2013. [Online]. Available: <https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0>. [Accessed: 03-Sep-2019]
- [8] S. Doherty, “The Impact of Translation Technologies on the Process and Product of Translation,” p. 23, 2016.
- [9] L. Ahrenberg, “Comparing Machine Translation and Human Translation: A Case Study,” in *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, Varna, Bulgaria, 2017, pp. 21–28, doi: 10.26615/978-954-452-042-7_003 [Online]. Available: https://doi.org/10.26615/978-954-452-042-7_003. [Accessed: 23-Jun-2020]
- [10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3 edition. Upper Saddle River: Pearson, 2009.
- [11] W. J. Hutchins, “Machine translation: history and general principles,” Asher (1994) p. 22-32.
- [12] W. J. Hutchins and H. L. Somers, *An Introduction to Machine Translation*. London: Academic Press, 1992.
- [13] B. Hettige and A. S. Karunananda, “Computational model of grammar for English to Sinhala Machine Translation,” in *2011 International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2011, pp. 26–31.
- [14] R. Zens, F. J. Och, and H. Ney, “Phrase-Based Statistical Machine Translation,” in *KI 2002: Advances in Artificial Intelligence*, 2002, pp. 18–32.
- [15] Y. Lu, P. Keung, F. Ladhak, V. Bhardwaj, S. Zhang, and J. Sun, “A neural interlingua for multilingual machine translation,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels, 2018, pp. 84–92 [Online]. Available: <http://www.aclweb.org/anthology/W18-6309>. [Accessed: 05-Feb-2019]

- [16] S. Nirenburg, “Knowledge-based machine translation,” *Mach. Transl.*, vol. 4, no. 1, pp. 5–24, Mar. 1989, doi: 10.1007/BF00367750.
- [17] H. Somers, “An Overview of EBMT,” in *Recent Advances in Example-Based Machine Translation*, M. Carl and A. Way, Eds. Dordrecht: Springer Netherlands, 2003, pp. 3–57 [Online]. Available: https://doi.org/10.1007/978-94-010-0181-6_1. [Accessed: 14-Jun-2020]
- [18] P. F. Brown *et al.*, “A Statistical Approach to Machine Translation,” *Comput. Linguist.*, vol. 16, no. 2, pp. 79–85, 1990.
- [19] P. Koehn, “Neural Machine Translation,” *ArXiv170907809 Cs*, Sep. 2017 [Online]. Available: <http://arxiv.org/abs/1709.07809>. [Accessed: 15-Sep-2019]
- [20] M. R. Costa-jussà and J. A. R. Fonollosa, “Latest trends in hybrid machine translation and its applications,” *Comput. Speech Lang.*, vol. 32, no. 1, pp. 3–10, Jul. 2015, doi: 10.1016/j.csl.2014.11.001.
- [21] S. Nirenburg, H. L. Somers, and Y. A. Wilks, “Treatment of Meaning in MT Systems,” in *Readings in Machine Translation*, MITP, 2003, pp. 281–293 [Online]. Available: <https://ieeexplore.ieee.org/document/6283755>. [Accessed: 23-Jun-2020]
- [22] J. Hajic, “Machine Translation of Very Close Languages,” in *ANLP*, 2000, doi: 10.3115/974147.974149.
- [23] W. J. Hutchins, “Machine Translation over fifty years,” *Hist. Épistémologie Lang.*, vol. 23, no. 1, pp. 7–31, 2001, doi: 10.3406/hel.2001.2815.
- [24] R. H. Richens, “Interlingual Machine Translation,” 1958, doi: 10.1093/comjnl/1.3.144.
- [25] “Machine Translation: Theoretical And Methodological Issues ONLINE FREE books in EPUB, TXT Sergei Nirenburg.” [Online]. Available: <http://banksmillersupply.com/machine-translation-theoretical-and-methodological-issues-us-pdf-allbooks-sergei-nirenburg.pdf>. [Accessed: 09-May-2019]
- [26] P. Koehn, *Statistical Machine Translation*, 1 edition. Cambridge ; New York: Cambridge University Press, 2009.
- [27] “Translation, Brains and the Computer | SpringerLink.” [Online]. Available: <https://link.springer.com/book/10.1007%2F978-3-319-76629-4>. [Accessed: 09-May-2019]
- [28] E. T. from the arXiv, “Human translators are still on top—for now,” *MIT Technology Review*. [Online]. Available: <https://www.technologyreview.com/s/611957/human-translators-are-still-on-top-for-now/>. [Accessed: 04-Feb-2019]
- [29] “MT Reaches Human Quality? Maybe, If You Squint Really Hard.” [Online]. Available: <http://www.common senseadvisory.com/Default.aspx?Contenttype=ArticleDetAD&tabID=63&Aid=48554&moduleId=390>. [Accessed: 23-Dec-2018]
- [30] “Translation procedures, strategies and methods.” [Online]. Available: <https://translationjournal.net/journal/41culture.htm>. [Accessed: 14-May-2019]
- [31] “Psycholinguistics/Parsing - Wikiversity.” [Online]. Available: <https://en.wikiversity.org/wiki/Psycholinguistics/Parsing>. [Accessed: 12-Mar-2019]

- [32] L. Frazier, "Sentence processing: A tutorial review," in *Attention and performance 12: The psychology of reading*, Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1987, pp. 559–586.
- [33] L. Frazier, "Constraint satisfaction as a theory of sentence processing," *J. Psycholinguist. Res.*, vol. 24, no. 6, pp. 437–468, Nov. 1995, doi: 10.1007/bf02143161.
- [34] "Garden Path Model And The Constraint-Based Model." [Online]. Available: <https://www.ukessays.com/essays/psychology/garden-path-model-and-the-constraint-based-model-psychology-essay.php>. [Accessed: 23-Jun-2020]
- [35] C. R. Huyck, "A psycholinguistic model of natural language parsing implemented in simulated neurons," *Cogn. Neurodyn.*, vol. 3, no. 4, pp. 317–330, Dec. 2009, doi: 10.1007/s11571-009-9080-6.
- [36] B. Hettige, A. S. Karunananda, and G. Rzevski, "MaSMT: A multi-agent system development framework for English-Sinhala machine translation," *Int. J. Comput. Linguist. Nat. Lang. Process. IJCLNLP*, vol. 2, no. 7, pp. 411–416, 2013.
- [37] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *CoRR*, vol. abs/1609.08144, 2016 [Online]. Available: <http://arxiv.org/abs/1609.08144>. [Accessed: 14-Jun-2019]
- [38] R. M. Weischedel, "Knowledge representation and natural language processing," *Proc. IEEE*, vol. 74, no. 7, pp. 905–920, Jul. 1986, doi: 10.1109/PROC.1986.13571.
- [39] T. Briscoe, "Introduction to Linguistics for Natural Language Processing," Computer Laboratory University of Cambridge, Michaelmas Term 2013.
- [40] Mertens, G. Strube, J. Dittmann, and H. Spada, "Human Sentence Processing : A Semantics-Oriented Parsing Approach," 2002.
- [41] "Human Sentence Processing Some Assumptions." [Online]. Available: http://www.l2f.inesc-id.pt/~abarreiro/openlogos-tutorial/human_sentence_processing_some_a.htm. [Accessed: 04-Feb-2019]
- [42] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-based Translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, Stroudsburg, PA, USA, 2003, pp. 48–54, doi: 10.3115/1073445.1073462 [Online]. Available: <https://doi.org/10.3115/1073445.1073462>. [Accessed: 27-Nov-2019]
- [43] R. Zens, F. J. Och, and H. Ney, "Phrase-Based Statistical Machine Translation," in *KI 2002: Advances in Artificial Intelligence*, Berlin, Heidelberg, 2002, pp. 18–32, doi: 10.1007/3-540-45751-8_2.
- [44] L. Osterhout, P. J. Holcomb, and D. A. Swinney, "Brain Potentials Elicited by Garden-Path Sentences: Evidence of the Application of Verb Information During Parsing," p. 18.
- [45] F. Ferreira, K. Christianson, and A. Hollingworth, "Misinterpretations of Garden-Path Sentences: Implications for Models of Sentence Processing and Reanalysis," *J. Psycholinguist. Res.*, vol. 30, no. 1, pp. 3–20, Jan. 2001, doi: 10.1023/A:1005290706460.

- [46] L. Frazier, "Constraint satisfaction as a theory of sentence processing," *J. Psycholinguist. Res.*, vol. 24, no. 6, pp. 437–468, Nov. 1995.
- [47] "Garden Path - an overview | ScienceDirect Topics." [Online]. Available: <https://www.sciencedirect.com/topics/psychology/garden-path>. [Accessed: 27-Nov-2019]
- [48] B. Hettige and A. S. Karunananda, "Existing Systems and Approaches for Machine Translation: A Review," in *Proceedings of the 8th Annual Sessions, Sri Lanka Association for Artificial Intelligence*, 2011 [Online]. Available: <http://slaai.lk/proc/2011/s1101.pdf>
- [49] "Interlingua in Google Translate | Daniel Stein - Way of the Word." [Online]. Available: <http://daniel-stein.com/node/269>. [Accessed: 07-Feb-2019]
- [50] J. Hutchins, "First steps in mechanical translation," MT Summit VI, 1997.
- [51] M. A. K. Halliday and E. Delavenay, "An Introduction to Machine Translation," *Mod. Lang. Rev.*, vol. 57, no. 1, p. 73, Jan. 1962, doi: 10.2307/3721978.
- [52] A. D. Booth, "Mechanical resolution of linguistic problems, Electronic Information handling. Washington, DC: Spartan Books, 1965.
- [53] Y. Bar-Hillel, "The present state of research on machine translation," *Am. Doc.*, vol. 2, no. 4, pp. 229–237, Oct. 1951, doi: 10.1002/asi.5090020408.
- [54] W. J. Hutchins, "Machine Translation: A Brief History," in *Concise History of the Language Sciences*, Elsevier, 1995, pp. 431–445 [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/B9780080425801500660>. [Accessed: 23-Jun-2020]
- [55] A. C. Reynolds, "The conference on mechanical translation held at M.I.T., June 17-20, 1952," p. 9.
- [56] A. Schenk, "Idioms in the Rosetta Machine Translation System," in *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, 1986 [Online]. Available: <https://www.aclweb.org/anthology/C86-1075>. [Accessed: 21-Jun-2020]
- [57] J. W. Perry, "Translation of Russian technical literature by machine", Machine Translation vol 2, 1995.
- [58] W. J. Hutchins, "Machine Translation over fifty years," *Hist. Épistémologie Lang.*, vol. 23, no. 1, pp. 7–31, 2001, doi: 10.3406/hel.2001.2815.
- [59] E. F. K. Koerner and R. E. Asher, *Concise History of the Language Sciences: From the Sumerians to the Cognitivists*. Elsevier, 2014.
- [60] "Free Online Translation | SYSTRAN Technologies." [Online]. Available: <http://www.systransoft.com/lp/free-online-translation/>. [Accessed: 23-Dec-2018]
- [61] "History of machine translation," *Wikipedia*. 24-Nov-2018 [Online]. Available: https://en.wikipedia.org/w/index.php?title=History_of_machine_translation&oldid=870326482. [Accessed: 05-Feb-2019]
- [62] S. AlAnsary, "Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas," p. 11.
- [63] R. H. Richens, "Interlingual Machine Translation," *Comput J*, vol. 1, pp. 144–147, 1958, doi: 10.1093/comjnl/1.3.144.
- [64] H. Uchida and M. Zhu, "Interlingua for multilingual machine translation, MT Summit IV 1993p 20-22.

- [65] A. Farghaly, “Arabic Machine Translation: A Developmental Perspective”, *International Journal on Information and Communication Technologies*, Vol. 3, No. 3, 2010 p3-10
- [66] X. Qi, H. Zhou, and H. Chen, “An interlingua-based Chinese-English MT system,” *J. Comput. Sci. Technol.*, vol. 17, no. 4, pp. 464–472, Jul. 2002, doi: 10.1007/BF02943286.
- [67] T. Modhiran, K. Kosawat, S. Klaithin, M. Boriboon, and T. Supnithi, *PARSIT TE: Online Thai-English Machine Translation*, MT Summit, 2005.
- [68] M. T. Wescoat, “Practical Instructions for Working with the Formalism of Lexical-Functional Grammar”, Summer Institute, Stanford University, 2005.
- [69] B. J. Dorr, “Interlingual machine translation A parameterized approach,” *Artif. Intell.*, vol. 63, no. 1, pp. 429–492, Oct. 1993, doi: 10.1016/0004-3702(93)90023-5.
- [70] S. Dave, J. Parikh, and P. Bhattacharyya, “Interlingua-based English–Hindi Machine Translation and Language Divergence,” *Mach. Transl.*, vol. 16, no. 4, pp. 251–304, Dec. 2001, doi: 10.1023/A:1021902704523.
- [71] “Universal Networking Language (UNL).” [Online]. Available: <http://language.worldofcomputing.net/unl/universal-networking-language-unl.html>. [Accessed: 14-Jun-2019]
- [72] M. Steyvers and J. B. Tenenbaum, “The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth,” *Cogn. Sci.*, vol. 29, no. 1, pp. 41–78, Jan. 2005, doi: 10.1207/s15516709cog2901_3.
- [73] H. S. Sreedeepta and S. M. Idicula, “Interlingua based Sanskrit-English machine translation,” in *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 2017, pp. 1–5, doi: 10.1109/ICCPCT.2017.8074251.
- [74] A. Bharati, V. Chaitanya, and R. Sangal, “Paninian framework and its application to Anusaraka,” *Sadhana*, vol. 19, no. 1, pp. 113–127, Feb. 1994, doi: 10.1007/BF02760393.
- [75] L. Bowker, “Computer-aided translation,” 2014, doi: 10.4324/9781315749129.ch4.
- [76] A. Taravella and A. O. Villeneuve, “Acknowledging the needs of computer-assisted translation tools users: the human perspective in human-machine translation,” 2013.
- [77] “OmegaT - The Free Translation Memory Tool - OmegaT,” *OmegaT - The Free Translation Memory Tool*. [Online]. Available: <http://omegat.org/>. [Accessed: 13-Jun-2019]
- [78] “OmegaT - The Free Translation Memory Tool - OmegaT.” [Online]. Available: <http://omegat.org/>. [Accessed: 08-Feb-2019]
- [79] “Translation software - memoQ.” [Online]. Available: <https://www.memoq.com/en/>. [Accessed: 08-Feb-2019]
- [80] “memoQ integration with machine translation (MT) systems.” [Online]. Available: <https://memoq.com/en/integration-with-machine-translation>. [Accessed: 13-Jun-2019]
- [81] H. Darbari, “Computer Assisted Translation System- An Indian Perspective,” MT Summit VII , 1999.

- [82] R. Sinha and A. Jain, “AnglaHindi: an English to Hindi machine-aided translation system,” p. 5.
- [83] R. M. K. Sinha, K. Sivaraman, A. Agrawal, R. Jain, R. Srivastava, and A. Jain, “ANGLABHARTI: a multilingual machine-aided translation project on translation from English to Indian languages,” in *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, 1995, vol. 2, pp. 1609–1614 vol.2, doi: 10.1109/ICSMC.1995.538002.
- [84] “MANTRA-RajBhasha.” [Online]. Available: <https://mantra-rajbhasha.rb-aii.in/>. [Accessed: 09-Feb-2019]
- [85] A. K. Joshi and Y. Schabes, “Tree-Adjoining Grammars,” in *Handbook of Formal Languages: Volume 3 Beyond Words*, G. Rozenberg and A. Salomaa, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 69–123 [Online]. Available: https://doi.org/10.1007/978-3-642-59126-6_2. [Accessed: 09-Feb-2019]
- [86] A. K. Joshi, “Mildly Context-Sensitive Grammars,” p. 4.
- [87] “Overall Salient Features of MANTRA System.” [Online]. Available: https://www.cdac.in/index.aspx?id=mc_mat_mantra_salient_features. [Accessed: 09-Feb-2019]
- [88] “ANUSAARAKA: OVERCOMING THE LANGUAGE BARRIER IN INDIA.” [Online]. Available: <https://ltrc.iiit.ac.in/Publications/anuvad.html>. [Accessed: 09-Feb-2019]
- [89] “Paninian Grammar Framework Applied to English.” [Online]. Available: https://ltrc.iiit.ac.in/Publications/pan_english.html. [Accessed: 09-Feb-2019]
- [90] S. Chaudhury, A. Rao, and D. M. Sharma, “Anusaaraka: An expert system based machine translation system,” in *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*, 2010, pp. 1–6, doi: 10.1109/NLPKE.2010.5587789.
- [91] RMM Shalini and B Hettige, “Dictionary-Based Machine Translation System for Pali to Sinhala,” in *Proceedings of the 13th Annual Sessions of Sri Sri Lanka Association for Artificial Intelligence*, Colombo, 2017.
- [92] J. Hajič, J. Hric, and V. Kuboň, “Machine translation of very close languages,” in *Proceedings of the sixth conference on Applied natural language processing*, Seattle, Washington, 2000, pp. 7–12, doi: 10.3115/974147.974149 [Online]. Available: <https://doi.org/10.3115/974147.974149>. [Accessed: 15-Jun-2020]
- [93] “Rule-based machine translation,” *Wikipedia*. 17-May-2019 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Rule-based_machine_translation&oldid=897553267. [Accessed: 14-Jun-2019]
- [94] “Apertium | A free/open-source machine translation platform.” [Online]. Available: <https://www.apertium.org/index.eng.html?dir=cat-por#translation>. [Accessed: 09-Feb-2019]
- [95] M. L. Forcada *et al.*, “Apertium: a free/open-source platform for rule-based machine translation,” *Mach. Transl.*, vol. 25, no. 2, pp. 127–144, Jun. 2011, doi: 10.1007/s10590-011-9090-0.
- [96] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep Transfer Learning with Joint Adaptation Networks,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, Sydney, NSW, Australia, 2017, pp. 2208–2217

- [Online]. Available: <http://dl.acm.org/citation.cfm?id=3305890.3305909>. [Accessed: 09-Feb-2019]
- [97] T. Izuha, A. Kumano, and Y. Kuroda, “Toshiba Rule-Based Machine Translation System at NTCIR-7 PAT MT,” in *NTCIR*, 2008.
- [98] E. Charniak, C. K. Riesbeck, D. V. McDermott, and J. R. Meehan, *Artificial Intelligence Programming*. Psychology Press, 1987.
- [99] B. Hettige and A. Karunananda, “On Demand Web Page Translation -BEES in action,” in *Proceeding of the sixth Annual Sessions*, Colombo, 2009, pp. 24–31 [Online]. Available: <http://www.slaai.lk/proc/2009/budditha.pdf>. [Accessed: 11-Mar-2016]
- [100] B. Hettige and A. S. Karunananda, “A Morphological analyzer to enable English to Sinhala Machine Translation,” in *Information and Automation, 2006. ICIA 2006. International Conference on*, 2006, pp. 21–26.
- [101] B. Hettige and A. S. Karunananda, “A Parser for Sinhala Language-First Step Towards English to Sinhala Machine Translation,” in *Industrial and Information Systems, First International Conference on*, 2006, pp. 583–587.
- [102] B. Hettige and A. S. Karunananda, “Computational model of grammar for English to Sinhala Machine Translation,” in *Advances in ICT for Emerging Regions (ICTer), 2011 International Conference on*, 2011, pp. 26–31.
- [103] B. Hettige and A. S. Karunananda, “Developing lexicon databases for English to Sinhala machine translation,” in *Industrial and Information Systems, 2007. ICIIS 2007. International Conference on*, 2007, pp. 215–220.
- [104] D. D. Silva *et al.*, “Sinhala to English Language Translator,” in *2008 4th International Conference on Information and Automation for Sustainability*, 2008, pp. 419–424, doi: 10.1109/ICIAFS.2008.4783983.
- [105] P. J. Antony. , “Machine Translation Approaches and Survey for Indian Languages,” in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 1, March 2013*, 2013 [Online]. Available: <https://www.aclweb.org/anthology/O13-2003>. [Accessed: 14-Jun-2019]
- [106] P. Desai, A. Sangodkar, and O. P. Damani, “A Domain-Restricted, Rule Based, English-Hindi Machine Translation System Based on Dependency Parsing,” *Proceedings of the 11th International Conference on Natural Language Processing*, 2014.
- [107] E. Universitat Politècnica de València, “Universitat Politècnica de València,” *Ing. Agua*, vol. 18, no. 1, p. ix, Sep. 2014, doi: 10.4995/ia.2014.3293.
- [108] S. Nirenburg, H. L. Somers, and Y. A. Wilks, Eds., “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle,” in *Readings in Machine Translation*, The MIT Press, 2003 [Online]. Available: <https://direct.mit.edu/books/book/2694/chapter/72867/a-framework-of-a-mechanical-translation-between>. [Accessed: 21-Jun-2020]
- [109] S. Kurohashi, T. Nakazawa, K. Alexis, and D. Kawahara, “Example-based machine translation pursuing fully structural NLP,” in *In Proc. of IWSLT’05*, 2005, pp. 207–212.
- [110] Ying Liu and Chengqing Zong, “Example-based Chinese-English MT,” in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE*

- Cat. No.04CH37583*), 2004, vol. 7, pp. 6093–6096 vol.7, doi: 10.1109/ICSMC.2004.1401354.
- [111] P. Unlee and P. Seresangtakul, “Thai to Isarn dialect machine translation using rule-based and example-based,” in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2016, pp. 1–5, doi: 10.1109/JCSSE.2016.7748892.
- [112] A. S. M. Kadhem and Y. R. Nasir, “English to Arabic Example-based Machine Translation System,” p. 17, 2015.
- [113] P. F. Brown *et al.*, “A STATISTICAL APPROACH TO MACHINE TRANSLATION,” *Comput. Linguist.*, vol. 1, no. 2, 1990 [Online]. Available: <http://aclweb.org/anthology/J/J90/J90-2002>. [Accessed: 09-Feb-2019]
- [114] “Bayes’ theorem,” *Wikipedia*. 06-Jun-2019 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bayes%27_theorem&oldid=900551935. [Accessed: 15-Jun-2019]
- [115] “Bayes’s theorem | Definition & Example | Britannica.” [Online]. Available: <https://www.britannica.com/topic/Bayess-theorem>. [Accessed: 23-Jun-2020]
- [116] “Moses - Main/HomePage.” [Online]. Available: <http://www.statmt.org/moses/>. [Accessed: 24-Dec-2014]
- [117] P. Koehn and Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation,” presented at the Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic, 2007.
- [118] “Babelfish.com.” [Online]. Available: <https://www.babelfish.com/>. [Accessed: 09-Feb-2019]
- [119] C. Kit and T. M. Wong, “Comparative Evaluation of Online Machine Translation Systems with Legal Texts,” *Law Libr. J.*, vol. 100, p. 23.
- [120] “Bing Microsoft Translator.” [Online]. Available: <https://www.bing.com/translator>. [Accessed: 09-Feb-2019]
- [121] “Google Translate.” [Online]. Available: <https://translate.google.com/?hl=en>. [Accessed: 23-Jun-2020]
- [122] Y. Wu *et al.*, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” Sep. 2016 [Online]. Available: <https://arxiv.org/abs/1609.08144>. [Accessed: 24-Jun-2017]
- [123] H. Ghasemi and M. Hashemian, “A Comparative Study of Google Translate Translations: An Error Analysis of English-to-Persian and Persian-to-English Translations,” *Engl. Lang. Teach.*, vol. 9, no. 3, p. 13, Jan. 2016, doi: 10.5539/elt.v9n3p13.
- [124] R. Pushpananda, R. Weerasinghe, and M. Niranjana, “Sinhala-Tamil Machine Translation: Towards better Translation Quality,” in *Proceedings of the Australasian Language Technology Association Workshop 2014*, Melbourne, Australia, 2014, pp. 129–133 [Online]. Available: <https://www.aclweb.org/anthology/U14-1018>. [Accessed: 09-Nov-2019]
- [125] S. Rajpirathap, S. Sheeyam, K. Umasuthan, and A. Chelvarajah, “Statistical Machine Translation System for Sinhala and Tamil Languages,” Apr. 2017

- [Online]. Available: <http://dl.lib.mrt.ac.lk/handle/123/12629>. [Accessed: 09-Nov-2019]
- [126] S. Goldwater and D. McClosky, “Improving statistical MT through morphological analysis,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 676–683 [Online]. Available: <http://dl.acm.org/citation.cfm?id=1220660>. [Accessed: 05-Sep-2017]
- [127] A. Ahmed and G. Hanneman, “Syntax-Based Statistical Machine Translation: A review,” *Comput. Linguist.*, p. 30.
- [128] S. Ranathunga, F. Farhath, U. Thayasivam, S. Jayasena, and G. Dias, “Si-Ta: Machine Translation of Sinhala and Tamil Official Documents,” in *2018 National Information Technology Conference (NITC)*, 2018, pp. 1–6, doi: 10.1109/NITC.2018.8550069.
- [129] P.-S. Huang, C. Wang, S. Huang, D. Zhou, and L. Deng, “TOWARDS NEURAL PHRASE-BASED MACHINE TRANSLATION,” p. 14, 2018.
- [130] *TensorFlow Neural Machine Translation Tutorial. Contribute to tensorflow/nmt development by creating an account on GitHub.* tensorflow, 2019 [Online]. Available: <https://github.com/tensorflow/nmt>. [Accessed: 10-Feb-2019]
- [131] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112 [Online]. Available: <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>. [Accessed: 10-Feb-2019]
- [132] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734 [Online]. Available: <http://www.aclweb.org/anthology/D14-1179>. [Accessed: 10-Feb-2019]
- [133] “Google brings offline neural machine translations for 59 languages to its Translate app,” *TechCrunch*. [Online]. Available: <http://social.techcrunch.com/2018/06/12/google-brings-offline-neural-machine-translation-for-59-languages-to-its-translate-app/>. [Accessed: 10-Feb-2019]
- [134] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, “OpenNMT: Open-Source Toolkit for Neural Machine Translation,” *ArXiv170102810 Cs*, Jan. 2017 [Online]. Available: <http://arxiv.org/abs/1701.02810>. [Accessed: 10-Feb-2019]
- [135] T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 1412–1421, doi: 10.18653/v1/D15-1166 [Online]. Available: <http://aclweb.org/anthology/D15-1166>. [Accessed: 10-Feb-2019]
- [136] P. Tennage *et al.*, “Neural machine translation for sinhala and tamil languages,” in *2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 189–192, doi: 10.1109/IALP.2017.8300576.

- [137] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” *ArXiv14090473 Cs Stat*, May 2016 [Online]. Available: <http://arxiv.org/abs/1409.0473>. [Accessed: 17-Jun-2020]
- [138] P. Koehn and R. Knowles, “Six Challenges for Neural Machine Translation,” in *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, 2017, pp. 28–39, doi: 10.18653/v1/W17-3204 [Online]. Available: <https://www.aclweb.org/anthology/W17-3204>. [Accessed: 09-Nov-2019]
- [139] P. Koehn and R. Knowles, “Six Challenges for Neural Machine Translation,” in *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, 2017, pp. 28–39, doi: 10.18653/v1/W17-3204.
- [140] K. Knight and S. K. Luk, “Building a Large-Scale Knowledge Base for Machine Translation,” *ArXivcmp-Lg9407029*, Jul. 1994 [Online]. Available: <http://arxiv.org/abs/cmp-lg/9407029>. [Accessed: 09-Feb-2019]
- [141] S. Nirenburg, V. Raskin, and A. Tucker, “ON KNOWLEDGE-BASED MACHINE TRANSLATION,” in *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, 1986 [Online]. Available: <http://aclweb.org/anthology/C/C86/C86-1148>. [Accessed: 09-Feb-2019]
- [142] “Knowledge-Based MT.” [Online]. Available: <https://www1.essex.ac.uk/linguistics/external/clmt/MTbook/HTML/node89.html>. [Accessed: 10-Feb-2019]
- [143] “KANT: Knowledge-Based Machine Translation | Carnegie Mellon University - Language Technologies Institute.” [Online]. Available: <https://www.lti.cs.cmu.edu/projects/machine-translation/kant-knowledge-based-machine-translation>. [Accessed: 09-Feb-2019]
- [144] A. Trujillo, “Transfer Machine Translation,” in *Translation Engines: Techniques for Machine Translation*, A. Trujillo, Ed. London: Springer, 1999, pp. 121–166 [Online]. Available: https://doi.org/10.1007/978-1-4471-0587-9_6. [Accessed: 21-Jun-2020]
- [145] M. Junczys-Dowmunt and R. Grundkiewicz, “Phrase-based Machine Translation is State-of-the-Art for Automatic Grammatical Error Correction,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, 2016, pp. 1546–1556 [Online]. Available: <https://aclweb.org/anthology/D16-1161>. [Accessed: 11-Feb-2019]
- [146] P.-S. Huang, C. Wang, S. Huang, D. Zhou, and L. Deng, “TOWARDS NEURAL PHRASE-BASED MACHINE TRANSLATION,” p. 14, 2018.
- [147] D. Ye and M. Zhang, “A Self-Adaptive Sleep/Wake-Up Scheduling Approach for Wireless Sensor Networks,” *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 979–992, Mar. 2018, doi: 10.1109/TCYB.2017.2669996.
- [148] M. Post, Y. Cao, and G. Kumar, “Joshua 6: A phrase-based and hierarchical statistical machine translation system,” *Prague Bull. Math. Linguist.*, vol. 104, no. 1, pp. 5–16, Oct. 2015, doi: 10.1515/pralin-2015-0009.
- [149] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation,” in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-*

- Volume 1*, 2003, pp. 48–54 [Online]. Available: <http://dl.acm.org/citation.cfm?id=1073462>. [Accessed: 22-Feb-2015]
- [150] N. Chatterjee and S. Gupta, “Efficient Phrase Table pruning for Hindi to English machine translation through syntactic and marker-based filtering and hybrid similarity measurement,” *Nat. Lang. Eng.*, vol. 25, no. 1, pp. 171–210, Jan. 2019, doi: 10.1017/S1351324918000360.
- [151] N. R. Prabhugaonkar, A. S. Nagvenkar, D. Kanojia, J. Pawar, P. Bhattacharyya, and M. Shrivastava, “PanchBhoota: Hierarchical Phrase Based Machine Translation Systems for Five Indian Languages,” p. 6.
- [152] S. P. Singh, A. Kumar, P. Sahu, and P. Verma, “Syntax based machine translation using blended methodology,” in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 242–247, doi: 10.1109/NGCT.2016.7877422.
- [153] P. Williams, R. Sennrich, M. Post, and P. Koehn, *Syntax-based statistical machine translation*, Computational Linguistics, 2016, p 893-896 2016.
- [154] K. Yamada and K. Knight, “A Syntax-based Statistical Translation Model,” in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001, pp. 523–530, doi: 10.3115/1073012.1073079 [Online]. Available: <https://www.aclweb.org/anthology/P01-1067>. [Accessed: 21-Jun-2020]
- [155] I. Minakov, G. Rzevski, P. Skobelev, and S. Volman, “Creating Contract Templates for Car Insurance Using Multi-agent Based Text Understanding and Clustering,” in *Holonic and Multi-Agent Systems for Manufacturing*, Berlin, Heidelberg, 2007, pp. 361–370, doi: 10.1007/978-3-540-74481-8_34.
- [156] M.-H. Stefanini and Y. Demazeau, “TALISMAN: A multi-agent system for natural language processing,” in *Advances in Artificial Intelligence*, Berlin, Heidelberg, 1995, pp. 312–322, doi: 10.1007/BFb0034824.
- [157] M. M. Aref, “A multi-agent system for natural language understanding,” in *IEMC '03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No.03CH37502)*, 2003, pp. 36–40, doi: 10.1109/KIMAS.2003.1245018.
- [158] C. Shi, T. Ishida, and D. Lin, “Translation Agent: A New Metaphor for Machine Translation,” *New Gener. Comput.*, vol. 32, no. 2, pp. 163–186, Apr. 2014, doi: 10.1007/s00354-014-0204-0.
- [159] T. Bi, H. Xiong, Z. He, H. Wu, and H. Wang, “Multi-agent Learning for Neural Machine Translation,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 856–865, doi: 10.18653/v1/D19-1079 [Online]. Available: <https://www.aclweb.org/anthology/D19-1079>. [Accessed: 16-Jun-2020]
- [160] K. Simov and P. Osenova, “A Hybrid Approach for Deep Machine Translation,” in *Proceedings of the 2nd Deep Machine Translation Workshop*, Praha, Czechia, 2016, pp. 21–28 [Online]. Available: <http://www.aclweb.org/anthology/W16-6403>. [Accessed: 11-Feb-2019]
- [161] “WordNet | A Lexical Database for English.” [Online]. Available: <https://wordnet.princeton.edu/>. [Accessed: 11-Feb-2019]

- [162] H. Hoang and P. Koehn, “Design of the Moses Decoder for Statistical Machine Translation,” in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, Columbus, Ohio, 2008, pp. 58–65 [Online]. Available: <http://www.aclweb.org/anthology/W/W08/W08-0510>. [Accessed: 11-Feb-2019]
- [163] O. Dhariya, S. Malviya, and U. S. Tiwary, “A hybrid approach for Hindi-English machine translation,” in *2017 International Conference on Information Networking (ICOIN)*, 2017, pp. 389–394, doi: 10.1109/ICOIN.2017.7899465.
- [164] N. de Silva, “Survey on Publicly Available Sinhala Natural Language Processing Tools and Research,” *ArXiv190602358 Cs*, Jan. 2020 [Online]. Available: <http://arxiv.org/abs/1906.02358>. [Accessed: 14-Jun-2020]
- [165] “Downloads | Language Technology Research Lab.” [Online]. Available: <http://ltrl.ucsc.lk/download-3/>. [Accessed: 12-Feb-2019]
- [166] R. Weerasinghe, D. Herath, and V. Welgama, “Corpus-based Sinhala Lexicon,” in *Proceedings of the 7th Workshop on Asian Language Resources*, Suntec, Singapore, 2009, pp. 17–23 [Online]. Available: <http://www.aclweb.org/anthology/W/W09/W09-3403>. [Accessed: 12-Feb-2019]
- [167] V. Welgama, D. L. Herath, C. Liyanage, N. Udalamatta, R. Weerasinghe, and T. Jayawardhane, “Towards a Sinhala Wordnet,” p. 5.
- [168] T. Nadungodage, R. Weerasinghe, and M. Niranjana, “Speaker Adaptation Applied to Sinhala Speech Recognition,” p. 13.
- [169] R. Pushpananda, R. Weerasinghe, and M. Niranjana, “Sinhala-Tamil Machine Translation: Towards better Translation Quality,” in *Proceedings of the Australasian Language Technology Association Workshop 2014*, Melbourne, Australia, 2014, pp. 129–133 [Online]. Available: <https://www.aclweb.org/anthology/U14-1018>. [Accessed: 28-Nov-2019]
- [170] “SinMin - Sinhala Corpus Project,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/306400561_SinMin_-_Sinhala_Corpus_Project. [Accessed: 12-Feb-2019]
- [171] S. Ranathunga, F. Farhath, U. Thayasivam, S. Jayasena, and G. Dias, “Si-Ta: Machine Translation of Sinhala and Tamil Official Documents,” in *2018 National Information Technology Conference (NITC)*, 2018, pp. 1–6, doi: 10.1109/NITC.2018.8550069.
- [172] S. Fernando, S. Ranathunga, S. Jayasena, and G. Dias, “Comprehensive Part-Of-Speech Tag Set and SVM based POS Tagger for Sinhala,” in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, Osaka, Japan, 2016, pp. 173–182 [Online]. Available: <https://www.aclweb.org/anthology/W16-3718>. [Accessed: 17-Jun-2020]
- [173] “(PDF) EnSiTip: A Tool to Unlock the English Web,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/268982590_EnSiTip_A_Tool_to_Unlock_the_English_Web. [Accessed: 04-Feb-2019]
- [174] W. Viraj, W. Ruvan, and M. Niranjana, “Defining the Gold Standard Definitions for the Morphology of Sinhala Words,” *Res. Comput. Sci.*, vol. 90, no. 1, pp. 163–171, Dec. 2015, doi: 10.13053/rcs-90-1-12.

- [175] R. Weerasinghe, “A Statistical Machine Translation Approach to Sinhala-Tamil Language Translation,” *ICT Enabled Soc.*, p. 136, 2003.
- [176] R. Pushpananda, R. Weerasinghe, and M. Niranjana, “Sinhala-Tamil Machine Translation: Towards better Translation Quality,” in *Proceedings of the Australasian Language Technology Association Workshop 2014*, Melbourne, Australia, 2014, pp. 129–133 [Online]. Available: <http://> [Accessed: 12-Feb-2019]
- [177] N. V. C. Vithanage, “English to Sinhala Intelligent Translator for Weather forecasting domain,” BIT degree, University of Colombo, Sri Lanka, Colombo, 2003.
- [178] B. T. L. Fernando, K. G. B. Gamage, K. T. S. Kasthuriarachchi, D. C. Jayasinghe, D. Chandrasena, and K. Pulasinghe, “English to Sinhala language Translator using Artificial Neural Networks,” *PSLIIT Vol2 SLIIT*, pp. 42–45, 2008.
- [179] L. Wijerathna *et al.*, “A Translator from Sinhala to English and English to Sinhala (SEES),” in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, 2012, pp. 14–18, doi: 10.1109/ICTer.2012.6421408.
- [180] B. Hettige, “A Computational grammar of Sinhala for English-Sinhala machine translation,” M.Phil Thesis, University of Moratuwa, Sri Lanka, Moratuwa, 2011 [Online]. Available: <http://dl.lib.mrt.ac.lk/handle/123/890>. [Accessed: 11-Mar-2016]
- [181] B. Hettige and A. Karunananda, “Theoretical based approach to English to Sinhala machine translation,” in *2009 International Conference on Industrial and Information Systems (ICIIS)*, 2009, pp. 380–385, doi: 10.1109/ICIINFS.2009.5429832.
- [182] B. Hettige and A. S. Karunananda, “Swarm intelligence of BEES for machine translation,” in *ITRU Research Symposium 2009*, Moratuwa, 2009 [Online]. Available: <http://dl.lib.mrt.ac.lk/handle/123/8409>. [Accessed: 12-Jul-2014]
- [183] W. Aroonmanakun, “Thoughts on Word and Sentence Segmentation in Thai,” p. 6.
- [184] N. Xue and Y. Yang, “Chinese sentence segmentation as comma classification,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, 2011, pp. 631–635 [Online]. Available: <http://www.aclweb.org/anthology/P11-2111>. [Accessed: 14-Feb-2019]
- [185] “Automatic Segmentation of Separately Pronounced Sinhala words into Syllables,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/265685421_Automatic_Segmentation_of_Separately_Pronounced_Sinhala_words_into_Syllables. [Accessed: 14-Feb-2019]
- [186] “Approaches to line breaking.” [Online]. Available: <http://w3c.github.io/i18n-drafts/articles/typography/linebreak.en>. [Accessed: 14-Feb-2019]
- [187] “English verb conjugation: past tense, participle, present perfect, past perfect | Reverso Conjugator.” [Online]. Available: <http://conjugator.reverso.net/conjugation-english.html>. [Accessed: 14-Feb-2019]

- [188] “Identify the tenses,” *English Grammar*, 02-May-2016. [Online]. Available: <https://www.englishgrammar.org/identify-tenses-2/>. [Accessed: 14-Feb-2019]
- [189] P. C. Wren and H. Martin, *High School English Grammar and Composition*, Revised edition. New Delhi: S Chand & Co Ltd, 1995.
- [190] A. M. Gunasekara, *A Comprehensive Grammar of the Sinhalese Language*. Asian Educational Services, 1999.
- [191] R. Boukobza and A. Rappoport, “Multi-Word Expression Identification Using Sentence Surface Features,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009, pp. 468–477 [Online]. Available: <http://www.aclweb.org/anthology/D/D09/D09-1049>. [Accessed: 14-Feb-2019]
- [192] “Multiword Expressions - ACL Wiki.” [Online]. Available: https://aclweb.org/aclwiki/Multiword_Expressions. [Accessed: 14-Feb-2019]
- [193] “OOV - Wiktionary.” [Online]. Available: <https://en.wiktionary.org/wiki/OOV>. [Accessed: 14-Feb-2019]
- [194] N. Habash, “Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation,” in *Proceedings of ACL-08: HLT, Short Papers*, Columbus, Ohio, 2008, pp. 57–60 [Online]. Available: <http://www.aclweb.org/anthology/P/P08/P08-2015>. [Accessed: 14-Feb-2019]
- [195] “How Many Mother Tongue Languages Are There?,” *Day Translations Blog*, 14-Jan-2018. [Online]. Available: <https://www.daytranslations.com/blog/2018/01/how-many-mother-tongue-languages-are-there-10529/>. [Accessed: 14-Feb-2019]
- [196] “Inflection,” *Wikipedia*. 15-Feb-2019 [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Inflection&oldid=883461766>. [Accessed: 28-Feb-2019]
- [197] “Words in English: Latin and Greek Morphology.” [Online]. Available: <http://www.ruf.rice.edu/~kemmer/Words/classmorph.html>. [Accessed: 28-Feb-2019]
- [198] “The morphology of -ly and the categorial status of ‘adverbs’ in English,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/259425996_The_morphology_of_-ly_and_the_categorial_status_of_'adverbs'_in_English. [Accessed: 08-Mar-2019]
- [199] “Basic English Sentence Structures - Sentence Types.” [Online]. Available: <https://www.scientificpsychic.com/grammar/enggram2.html>. [Accessed: 28-Feb-2019]
- [200] “Reviews: Levels of language.” [Online]. Available: https://www.uni-due.de/SHE/REV_Levels_Chart.htm. [Accessed: 08-Mar-2019]
- [201] L. Hennig, T. Strecker, S. Narr, E. W. De Luca, and S. Albayrak, “Identifying Sentence-Level Semantic Content Units with Topic Models,” in *2010 Workshops on Database and Expert Systems Applications*, Bilbao, TBD, Spain, 2010, pp. 59–63, doi: 10.1109/DEXA.2010.33 [Online]. Available: <http://ieeexplore.ieee.org/document/5592003/>. [Accessed: 08-Mar-2019]

- [202] W. Zadrozny and K. Jensen, “Semantics of Paragraphs,” *Comput. Linguist.*, vol. 1, no. 2, 1991 [Online]. Available: <http://aclweb.org/anthology/J/J91/J91-2003>. [Accessed: 08-Mar-2019]
- [203] “The alphabet ~ භ්‍යවේශ - Wikibooks, open books for an open world.” [Online]. Available: <https://en.wikibooks.org/wiki/Sinhala/1.2>. [Accessed: 09-Mar-2019]
- [204] “TDIL-DC :Morphological analyzer.” [Online]. Available: http://tdil-dc.in/index.php?option=com_vertical&parentid=60&lang=en. [Accessed: 10-Mar-2019]
- [205] B. Hettige and A. S. Karunananda, “A Morphological Analyzer to Enable English to Sinhala Machine Translation,” in *International Conference on Information and Automation, 2006. ICIA 2006*, 2006, pp. 21–26, doi: 10.1109/ICINFA.2006.374146.
- [206] S. Lushanthan, A. R. Weerasinghe, and D. L. Herath, “Morphological analyzer and generator for Tamil Language,” in *2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2014, pp. 190–196, doi: 10.1109/ICTER.2014.7083900.
- [207] K. R. Beesley and L. Karttunen, “Finite-State Morphology: Xerox Tools and Techniques ——— Pre-Publication Review Copy ——— Do Not Quote, Copy or Redistribute,” p. 690.
- [208] V. Goyal and G. S. Lehal, “Hindi Morphological Analyzer and Generator,” in *2008 First International Conference on Emerging Trends in Engineering and Technology*, 2008, pp. 1156–1159, doi: 10.1109/ICETET.2008.11.
- [209] “(PDF) Hindi Morphological Analyzer and Generator,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/232654434_Hindi_Morphological_Analyzer_and_Generator. [Accessed: 09-Mar-2019]
- [210] M. Bapat, H. Gune, and P. Bhattacharyya, “A Paradigm-Based Finite State Morphological Analyzer for Marathi,” in *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, Beijing, China, 2010, pp. 26–34 [Online]. Available: <http://www.aclweb.org/anthology/W10-3604>. [Accessed: 10-Mar-2019]
- [211] D. Alfter, “Analyzer and generator for Pali,” *ArXiv151001570 Cs*, Oct. 2015 [Online]. Available: <http://arxiv.org/abs/1510.01570>. [Accessed: 10-Mar-2019]
- [212] “A Rule based Kannada Morphological Analyzer and Generator using Finite State Transducer | Request PDF,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/264843084_A_Rule_based_Kannada_Morphological_Analyzer_and_Generator_using_Finite_State_Transducer. [Accessed: 10-Mar-2019]
- [213] A. Bharati, V. Chaitanya, and R. Sangal, “Panel: Computational Linguistics in India: An Overview,” in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 2000, pp. 1–2, doi: 10.3115/1075218.1075295 [Online]. Available: <http://www.aclweb.org/anthology/P00-1077>. [Accessed: 09-Mar-2019]

- [214] “Generic morphological analysis shell,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/228771551_Generic_morphological_analysis_shell. [Accessed: 09-Mar-2019]
- [215] R. N. Horspool, “Recursive ascent-descent parsers,” in *Compiler Compilers*, vol. 477, D. Hammer, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991, pp. 1–10 [Online]. Available: http://link.springer.com/10.1007/3-540-53669-8_70. [Accessed: 23-Jun-2020]
- [216] S. M. Shieber, Y. Schabes, and F. C. N. Pereira, “Principles and implementation of deductive parsing,” *J. Log. Program.*, vol. 24, no. 1–2, pp. 3–36, Jul. 1995, doi: 10.1016/0743-1066(95)00035-I.
- [217] “Natural Language Toolkit — NLTK 3.4.5 documentation.” [Online]. Available: <https://www.nltk.org/>. [Accessed: 25-Nov-2019]
- [218] “Apache OpenNLP.” [Online]. Available: <https://opennlp.apache.org/>. [Accessed: 25-Nov-2019]
- [219] “JavaCC - The Java Parser Generator.” [Online]. Available: <https://javacc.org/>. [Accessed: 25-Nov-2019]
- [220] “The Stanford Natural Language Processing Group.” [Online]. Available: <https://nlp.stanford.edu/software/lex-parser.shtml>. [Accessed: 25-Nov-2019]
- [221] Jing Ding, D. Berleant, Jun Xu, and A. W. Fulmer, “Extracting biochemical interactions from MEDLINE using a link grammar parser,” in *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp. 467–471, doi: 10.1109/TAI.2003.1250226.
- [222] “Enju - A fast, accurate, and deep parser for English.” [Online]. Available: <http://www.nactem.ac.uk/enju/>. [Accessed: 25-Nov-2019]
- [223] B. Hettige and A. S. Karunananda, “A Parser for Sinhala Language-First Step Towards English to Sinhala Machine Translation,” in *Industrial and Information Systems, First International Conference on*, 2006, pp. 583–587.
- [224] B. Hettige and A. S. Karunananda, “A Parser for Sinhala Language - First Step Towards English to Sinhala Machine Translation,” in *First International Conference on Industrial and Information Systems*, 2006, pp. 583–587, doi: 10.1109/ICIIS.2006.365795.
- [225] Biplav Sarma, Anup Kumar Barman, and Gauhati University, “A Comprehensive Survey of Noun Phrase Chunking in Natural Languages,” *Int. J. Eng. Res.*, vol. V4, no. 04, p. IJERTV4IS040854, Apr. 2015, doi: 10.17577/IJERTV4IS040854.
- [226] “Multi-Agent Systems: A survey.” [Online]. Available: https://www.researchgate.net/publication/324847369_Multi-Agent_Systems_A_survey. [Accessed: 30-Mar-2019]
- [227] “Ontologies - Introduction to ontologies and semantic web - tutorial.” [Online]. Available: <https://www.obitko.com/tutorials/ontologies-semantic-web/ontologies.html>. [Accessed: 18-Sep-2019]
- [228] “Intelligent Software Agents.” [Online]. Available: <https://www.cs.cmu.edu/~softagents/multi.html>. [Accessed: 19-Sep-2019]
- [229] P. Dasgupta, “A Peer-to-Peer System Architecture for Multi-Agent Collaboration,” in *Intelligent Systems Design and Applications*, 2003, pp. 483–492.

- [230] “Multiagent Systems.” [Online]. Available: <https://www.cs.cmu.edu/afs/cs/usr/pstone/public/papers/97MAS-survey/node2.html>. [Accessed: 19-Sep-2019]
- [231] G. Rzevski and P. Skobelev, *Managing Complexity*. Southampton Boston: WIT Press, 2014.
- [232] “Why Coding Multi-Agent Systems is Hard – Hacker Noon.” [Online]. Available: <https://hackernoon.com/why-coding-multi-agent-systems-is-hard-2064e93e29bb>. [Accessed: 30-Mar-2019]
- [233] “FIPA Agent Communication Language Specifications.” [Online]. Available: <http://www.fipa.org/repository/aclspecs.html>. [Accessed: 30-Mar-2019]
- [234] “KQML as an agent communication language.” [Online]. Available: <https://dl.acm.org/citation.cfm?id=191322>. [Accessed: 30-Mar-2019]
- [235] Y. Labrou, T. Finin, and Yun Peng, “Agent communication languages: the current landscape,” *IEEE Intell. Syst.*, vol. 14, no. 2, pp. 45–52, Mar. 1999, doi: 10.1109/5254.757631.
- [236] F. Bellifemine, A. Poggi, and G. Rimassa, “Developing multi-agent systems with a FIPA-compliant agent framework,” p. 26, 2001.
- [237] “Jade Site | Java Agent DEvelopment Framework.” [Online]. Available: <https://jade.tilab.com/>. [Accessed: 20-Sep-2019]
- [238] “The MadKit Agent Platform Architecture,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/2646635_The_MadKit_Agent_Platform_Architecture. [Accessed: 20-Sep-2019]
- [239] “MaDKit.” [Online]. Available: <http://www.madkit.net/madkit/>. [Accessed: 31-Mar-2019]
- [240] O. Gutknecht and J. Ferber, “The MADKIT Agent Platform Architecture,” in *Revised Papers from the International Workshop on Infrastructure for Multi-Agent Systems: Infrastructure for Agents, Multi-Agent Systems, and Scalable Multi-Agent Systems*, London, UK, UK, 2001, pp. 48–55 [Online]. Available: <http://dl.acm.org/citation.cfm?id=646675.701833>. [Accessed: 31-Mar-2019]
- [241] “Gumroad.” [Online]. Available: https://gumroad.com/overlay_page. [Accessed: 31-Mar-2019]
- [242] “Python Agent DEvelopment framework — Pade 1.0 documentation.” [Online]. Available: <https://pade.readthedocs.io/en/latest/>. [Accessed: 21-Sep-2019]
- [243] G. Radhakrishnan and S. KI, “COMPARATIVE STUDY OF JADE AND SPADE MULTI AGENT SYSTEM.,” *Int. J. Adv. Res.*, vol. 6, no. 11, pp. 1035–1042.
- [244] M. E. Gregori, *ABSTRACT A Jabber-based Multi-Agent System Platform*, Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems, 2006
- [245] “Jason | a Java-based interpreter for an extended version of AgentSpeak.” [Online]. Available: <http://jason.sourceforge.net/wp/>. [Accessed: 23-Jun-2020]
- [246] R. H. Bordini and J. F. Hübner, “BDI Agent Programming in AgentSpeak Using Jason,” in *Computational Logic in Multi-Agent Systems*, Berlin, Heidelberg, 2006, pp. 143–164, doi: 10.1007/11750734_9.

- [247] “AgentBuilder.” [Online]. Available: <https://www.agentbuilder.com/>. [Accessed: 31-Mar-2019]
- [248] “SeSAm - Integrated Environment for Multi-Agent Simulation.” [Online]. Available: <http://www.simsesam.de/>. [Accessed: 31-Mar-2019]
- [249] F. Klügl and F. Puppe, “The Multi-Agent Simulation Environment SeSAm,” in *University Paderborn*, 1998.
- [250] H. Xu and S. M. Shatz, “ADK: An Agent Development Kit Based on a Formal Design Model for Multi-Agent Systems,” *Autom. Softw. Eng.*, vol. 10, no. 4, pp. 337–365, Oct. 2003, doi: 10.1023/A:1025859021913.
- [251] J. Ferber and O. Gutknecht, “A meta-model for the analysis and design of organizations in multi-agent systems,” in *Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160)*, 1998, pp. 128–135, doi: 10.1109/ICMAS.1998.699041.
- [252] “MaDKit.” [Online]. Available: <http://www.madkit.net/madkit/madkit.php>. [Accessed: 15-Aug-2017]
- [253] “MaSMT 3.0 Development Guide,” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/319101813_MaSMT_30_Development_Guide. [Accessed: 08-Dec-2018]
- [254] B. Hettige and A. S. Karunananda, “Octopus: A Multi Agent Chatbot,” Proceedings of 8th International Research Conference, KDU, 2015.
- [255] H. Jayarathna and B. Hettige, “AgriCom: A communication platform for agriculture sector,” in *Industrial and Information Systems (ICIIS), 2013 8th IEEE International Conference on*, 2013, pp. 439–444.
- [256] M. A. S. T. Goonatilleke, M. W. G. Jayampath, and B. Hettige, “Rice Express: A Communication Platform for Rice Production Industry,” in *Artificial Intelligence*, 2019, pp. 269–277.
- [257] L. Weerasinghe, B. Hettige, R. P. S. Kathriarachchi, and A. S. Karunananda, “Resource Sharing in Distributed Environment using Multi-agent Technology,” *Resource*, vol. 167, no. 5, 2017.
- [258] T. D. Samaranayake, W. P. J. Pamarathane, and B. Hettige, “Solution for event-planning using multi-agent technology,” in *2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2017, pp. 1–6, doi: 10.1109/ICTER.2017.8257805.
- [259] K. Christianson, A. Hollingworth, J. F. Halliwell, and F. Ferreira, “Thematic Roles Assigned along the Garden Path Linger,” *Cognit. Psychol.*, vol. 42, pp. 368–407, 2001, doi: 10.1006/cogp.2001.0752.
- [260] L. Frazier, “Constraint satisfaction as a theory of sentence processing,” *J. Psycholinguist. Res.*, vol. 24, no. 6, pp. 437–468, Nov. 1995, doi: 10.1007/BF02143161.
- [261] D. Yu, W. Wei, L. Jia, and B. Xu, “Confidence estimation for spoken language translation based on Round Trip Translation,” in *2010 7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 426–429, doi: 10.1109/ISCSLP.2010.5684855.
- [262] M. Kulathunga, “Madura English-Sinhala Dictionary - Online Language Translator.” [Online]. Available: <https://www.maduraonline.com/>. [Accessed: 04-Feb-2019]

- [263] “Bhasha Dictionary | Sinhala-English Dictionary.” [Online]. Available: <http://www.bhasha.lk/products/dictionary>. [Accessed: 04-Feb-2019]
- [264] J. Tomás, J. À. Mas, and F. Casacuberta, “A Quantitative Method for Machine Translation Evaluation,” in *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, Columbus, Ohio, 2003, pp. 27–34 [Online]. Available: <https://www.aclweb.org/anthology/W03-2804>. [Accessed: 21-Jun-2019]
- [265] M. Popović and H. Ney, “Towards Automatic Error Analysis of Machine Translation Output,” *Comput. Linguist.*, vol. 37, no. 4, pp. 657–688, 2011, doi: 10.1162/COLI_a_00072.
- [266] “Human Evaluation of Machine Translation,” 26-Jun-2016. [Online]. Available: <https://www.ebayinc.com/stories/blogs/tech/human-evaluation-of-machine-translation/>. [Accessed: 21-Jun-2019]
- [267] “Round Trip Translation Using PNMT Systems | SYSTRAN.” [Online]. Available: <https://blog.systransoft.com/round-trip-translation-no-more-entertainment-with-pnmt-systems/>. [Accessed: 23-Aug-2019]
- [268] “Round-trip translation - Semantic Scholar.” [Online]. Available: <https://www.semanticscholar.org/topic/Round-trip-translation/1058180>. [Accessed: 21-Jun-2019]
- [269] H. Somers, “Round-trip translation: what is it good for,” in *In proceedings of the Australasian Language Technology Workshop*, 2005, pp. 127–133.
- [270] M. Popović and H. Ney, “Word error rates: decomposition over Pos classes and applications for error analysis,” in *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, Prague, Czech Republic, 2007, pp. 48–55, doi: 10.3115/1626355.1626362 [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1626355.1626362>. [Accessed: 21-Jun-2019]
- [271] “Levenshtein Distance - an overview | ScienceDirect Topics.” [Online]. Available: <https://www.sciencedirect.com/topics/computer-science/levenshtein-distance>. [Accessed: 21-Jun-2019]
- [272] S. Rani and J. Singh, “Enhancing Levenshtein’s Edit Distance Algorithm for Evaluating Document Similarity,” in *Computing, Analytics and Networks*, 2018, pp. 72–80.
- [273] M. Thoma, “Word Error Rate Calculation,” *Martin Thoma*. [Online]. Available: <http://www.martin-thoma.de/word-error-rate-calculation/>. [Accessed: 08-Sep-2019]
- [274] L. Han, “Machine Translation Evaluation Resources and Methods: A Survey,” *ArXiv160504515 Cs*, Sep. 2018 [Online]. Available: <http://arxiv.org/abs/1605.04515>. [Accessed: 23-Jun-2020]
- [275] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A Study of Translation Edit Rate with Targeted Human Annotation,” p. 9.
- [276] M. Popovic, “Class error rates for evaluation of machine translation output,” *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 2012
- [277] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA,

- USA, 2002, pp. 311–318, doi: 10.3115/1073083.1073135 [Online]. Available: <https://doi.org/10.3115/1073083.1073135>. [Accessed: 21-Jun-2019]
- [278] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a Method for Automatic Evaluation of Machine Translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318, doi: 10.3115/1073083.1073135 [Online]. Available: <https://www.aclweb.org/anthology/P02-1040>. [Accessed: 21-Jun-2019]
- [279] B. Chen and C. Cherry, “A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU,” in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, 2014, pp. 362–367, doi: 10.3115/v1/W14-3346 [Online]. Available: <http://aclweb.org/anthology/W14-3346>. [Accessed: 08-Sep-2019]
- [280] “Index — NLTK 3.4.5 documentation.” [Online]. Available: <https://www.nltk.org/genindex.html>. [Accessed: 14-Sep-2019]
- [281] “donnabelldmello/nlp-bleu,” *GitHub*. [Online]. Available: <https://github.com/donnabelldmello/nlp-bleu>. [Accessed: 13-Sep-2019]
- [282] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” p. 8.
- [283] A. Gupta, S. Venkatapathy, and R. Sangal, “METEOR-Hindi : Automatic MT Evaluation Metric for Hindi as a Target Language,” 2010.
- [284] M. Denkowski and A. Lavie, “Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks,” p. 9.
- [285] E. Chatzikoumi, “How to evaluate machine translation: A review of automated and human metrics,” *Nat. Lang. Eng.*, vol. 26, no. 2, pp. 137–161, Mar. 2020, doi: 10.1017/S1351324919000469.
- [286] M. S. Maučec and G. Donaj, “Machine Translation and the Evaluation of Its Quality,” *Recent Trends Comput. Intell.*, Sep. 2019, doi: 10.5772/intechopen.89063. [Online]. Available: <https://www.intechopen.com/books/recent-trends-in-computational-intelligence/machine-translation-and-the-evaluation-of-its-quality>. [Accessed: 22-Jun-2020]
- [287] “Evaluation of machine translation,” *Wikipedia*. 09-Nov-2017 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Evaluation_of_machine_translation&oldid=809464440. [Accessed: 10-Nov-2017]
- [288] “A Comparatives Study of Machine Translation Evaluation Systems | July 2016 | Translation Journal.” [Online]. Available: <https://translationjournal.net/July-2016/a-comparatives-study-of-machine-translation-evaluation-systems.html>. [Accessed: 22-Jun-2020]
- [289] P. Koehn and C. Monz, “Manual and automatic evaluation of machine translation between European languages,” in *Proceedings of the Workshop on Statistical Machine Translation - StatMT '06*, New York City, New York, 2006, p. 102, doi: 10.3115/1654650.1654666 [Online]. Available:

- <http://portal.acm.org/citation.cfm?doid=1654650.1654666>. [Accessed: 21-Jun-2019]
- [290] H. Li, “Adequacy-Fluency Metrics (AM-FM) for Machine Translation (MT) Evaluation,” p. 45.
- [291] G. M. Sullivan and A. R. Artino, “Analyzing and Interpreting Data From Likert-Type Scales,” *J. Grad. Med. Educ.*, vol. 5, no. 4, pp. 541–542, Dec. 2013, doi: 10.4300/JGME-5-4-18.
- [292] J. Sim and C. C. Wright, “The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements,” *Phys. Ther.*, vol. 85, no. 3, pp. 257–268, Mar. 2005, doi: 10.1093/ptj/85.3.257.
- [293] “Kappa Statistics - an overview | ScienceDirect Topics.” [Online]. Available: <https://www.sciencedirect.com/topics/medicine-and-dentistry/kappa-statistics>. [Accessed: 03-Aug-2019]
- [294] “Cohen’s Kappa | Real Statistics Using Excel.” [Online]. Available: <http://www.real-statistics.com/reliability/interrater-reliability/cohens-kappa/>. [Accessed: 15-Sep-2019]
- [295] M. Martindale and M. Carpuat, “Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT,” in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, Boston, MA, 2018, pp. 13–25 [Online]. Available: <https://www.aclweb.org/anthology/W18-1803>. [Accessed: 24-Aug-2019]
- [296] L. Puka, “Kendall’s Tau,” in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer, 2011, pp. 713–715 [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_324. [Accessed: 23-Jun-2020]
- [297] “How good is Google translate? The most accurate language pairs.” [Online]. Available: <https://www.betranslated.com/blog/how-good-is-google-translate/>. [Accessed: 09-Jul-2020]
- [298] “5 Reasons Not to Rely on Google Translate,” *Clear Words Translations*, 18-Oct-2019. [Online]. Available: <http://clearwordstranslations.com/5-reasons-not-to-rely-on-google-translate/>. [Accessed: 09-Jul-2020]
- [299] “The Pros and Cons of Google Translate,” *Language Connections*. [Online]. Available: <https://www.languageconnections.com/blog/the-pros-cons-of-google-translate/>. [Accessed: 11-Jul-2020]
- [300] “Google Translate accuracy – why it’s such a mixed bag,” *PacTranz*, 26-Aug-2014. [Online]. Available: <https://www.pactranz.com/google-translate-accuracy-issues/>. [Accessed: 11-Jul-2020]
- [301] X. Chen, S. Acosta, and A. E. Barry, “Evaluating the Accuracy of Google Translate for Diabetes Education Material,” *JMIR Diabetes*, vol. 1, no. 1, Jun. 2016, doi: 10.2196/diabetes.5848. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6238856/>. [Accessed: 11-Jul-2020]
- [302] *Sentences, English Lingistics for Language teaching*. Department of Distance Education, NIE.

Appendix A: Translation Summary with Agents' Communications

The following sample shows the translation summary .

run:

```
[genrate.701@siph]           [OK]
[analysis.101@ema]          [OK]
[analysis.102@ema]          [OK]
[analysis.103@ema]          [OK]
[analysis.104@ema]          [OK]
[analysis.105@ema]          [OK]
[analysis.106@ema]          [OK]
[analysis.107@ema]          [OK]
[analysis.108@ema]          [OK]
[analysis.109@ema]          [OK]
[analysis.201@esa]          [OK]
[manager.1@ensimas]         [MP-OK]
#11 agents are ready for translation
#input: the good boy and beautiful girl read a good book at the school
```

Word list

```
-----
1- the--
2- good--
3- boy--
4- and--
5- beautiful--
6- girl--
7- read--
8- a--
9- good--
10- book--
11- at--
12- the--
13- school--
[manager.1@ensimas( SEND TO GROUP ema)] - - - analyze-eng-morp  >>
[analysis.100@ema]
[analysis.109@ema] - - - update-eng-morp >> [manager.1@ensimas]
[analysis.107@ema] - - - update-eng-morp >> [manager.1@ensimas]
[analysis.102@ema] - - - update-eng-morp >> [manager.1@ensimas]
[analysis.108@ema] - - - update-eng-morp >> [manager.1@ensimas]
[analysis.103@ema] - - - update-eng-morp >> [manager.1@ensimas]
EMA Out | #[analysis.104@ema] - - - update-eng-morp >> [manager.1@ensimas]
[analysis.105@ema] - - - update-eng-morp >> [manager.1@ensimas]
```

###[analysis.106@ema] - - - update-eng-morp >> [manager.1@ensimas]
#[analysis.101@ema] - - - update-eng-morp >> [manager.1@ensimas]

..... OK
English word Morphology list

- : 10- book-40000249 (VBP)
: 13- school-40001732 (VBP)
: 2- good-6000001 (JJX)
: 9- good-6000001 (JJX)
: 2- good-70002755 (JJX)
: 5- beautiful-70000557 (JJX)
: 7- read-70005228 (JJX)
: 9- good-70002755 (JJX)
: 2- good-20005131 (NWS)
: 3- boy-20001393 (NWS)
: 6- girl-20005051 (NWS)
: 9- good-20005131 (NWS)
: 10- book-20001327 (NWS)
: 13- school-20009600 (NWS)
: 1- the-60001115 (DET)
: 2- good-60000463 (RBX)
: 4- and-60000117 (CON)
: 8- a-60001116 (DET)
: 9- good-60000463 (RBX)
: 11- at-60001121 (PRP)
: 12- the-60001115 (DET)
: 7- read-30002107 (VBP)
: 7- read-30002107 (VBD)
: 7- read-30002107 (VBN)

++++
English Morphological Analysis Completed

- ++++
1- the-- DET
2- good-- JJX JJX NWS RBX
3- boy-- NWS
4- and-- CON
5- beautiful-- JJX
6- girl-- NWS
7- read-- JJX VBP VBD VBN
8- a-- DET
9- good-- JJX JJX NWS RBX
10- book-- VBP NWS
11- at-- PRP
12- the-- DET
13- school-- VBP NWS

Ready for Syntax Analysis

[manager.1@ensimas] - - - analyze-eng-syn >> [analysis.201@esa]
Available EP : NP(3-DET,JJX,NWS-3-NWS)- 1 -3
Available EP : NP(3-DET,JJX,NWS-3-NWS)- 8 -10
Available EP : NP(4-DET,NWS-2-NWS)- 1 -2
Available EP : NP(4-DET,NWS-2-NWS)- 8 -9
Available EP : NP(4-DET,NWS-2-NWS)- 12 -13
Available EP : CN(35-CON-1-CON)- 4 -4
Available EP : PR(36-PRP-1-PRP)- 11 -11
Available EP : PP(40-PRP,DET,NWS-3-NWS)- 11 -13
Available EP : VP(45-VBP-1-VBP)- 7 -7
Available EP : VP(45-VBP-1-VBP)- 10 -10
Available EP : VP(45-VBP-1-VBP)- 13 -13
Available EP : VP(46-VBD-1-VBD)- 7 -7
Available EP : VE(94-VBN-1-VBN)- 7 -7
Available EP : NP(100-NWS-1-NWS)- 2 -2
Available EP : NP(100-NWS-1-NWS)- 3 -3
Available EP : NP(100-NWS-1-NWS)- 6 -6
Available EP : NP(100-NWS-1-NWS)- 9 -9
Available EP : NP(100-NWS-1-NWS)- 10 -10
Available EP : NP(100-NWS-1-NWS)- 13 -13
Available EP : AP(104-RBX-1-RBX)- 2 -2
Available EP : AP(104-RBX-1-RBX)- 9 -9
Available EP : NP(108-JJX,NWS-2-NWS)- 2 -3
Available EP : NP(108-JJX,NWS-2-NWS)- 5 -6
Available EP : NP(108-JJX,NWS-2-NWS)- 9 -10

Total number of phrase available : 24

----- English Phrase list

0: NP-(1, 3-DET,JJX,NWS) NOS, NWS
0: NP-(1, 2-DET,NWS) NOS, NWS
0: NP-(2, 2-NWS) NOS, NWS
0: AP-(2, 2-RBX) AVP, RBX
0: NP-(2, 3-JJX,NWS) NOS, NWS
0: NP-(3, 3-NWS) NOS, NWS
0: CN-(4, 4-CON) CON, CON
0: NP-(5, 6-JJX,NWS) NOS, NWS
0: NP-(6, 6-NWS) NOS, NWS
0: VP-(7, 7-VBP) ACT-SPT, VBP
0: VP-(7, 7-VBD) ACT-PAT, VBD
0: VE-(7, 7-VBN) PRE-PAT, VBN
0: NP-(8, 10-DET,JJX,NWS) NOS, NWS
0: NP-(8, 9-DET,NWS) NOS, NWS

0: NP-(9, 9-NWS) NOS, NWS
0: AP-(9, 9-RBX) AVP, RBX
0: NP-(9, 10-JJX,NWS) NOS, NWS
0: VP-(10, 10-VBP) ACT-SPT, VBP
0: NP-(10, 10-NWS) NOS, NWS
0: PR-(11, 11-PRP) PRP, PRP
0: PP-(11, 13-PRP,DET,NWS) NOS, NWS
0: NP-(12, 13-DET,NWS) NOS, NWS
0: VP-(13, 13-VBP) ACT-SPT, VBP
0: NP-(13, 13-NWS) NOS, NWS

word count 1
MAX EP 3 - 0
word count 4
MAX EP 4 - 6
word count 5
MAX EP 6 - 7
word count 7
MAX EP 7 - 9
word count 8
MAX EP 10 - 12
word count 11
MAX EP 11 - 19
MAX EP 13 - 20
English Phrase list

1: NP-(1, 3-DET,JJX,NWS) NOS, NWS
2: CN-(4, 4-CON) CON, CON
3: NP-(5, 6-JJX,NWS) NOS, NWS
4: VP-(7, 7-VBP) ACT-SPT, VBP
5: NP-(8, 10-DET,JJX,NWS) NOS, NWS
6: PP-(11, 13-PRP,DET,NWS) NOS, NWS
SUB 3- none(TS)

OBJ 5- none

FVB4- none

[analysis.201@esa] - - - ready-sin-ph-generation >> [manager.1@ensimas]

Ready for Sinhala phrase generation

Generated Word based Ontology for input sentence

book(10-VBP) කලින් වෙන් කරනවා-veb වෙන් කරනවා-veb
school(13-VBP) දියුණු කරනවා-veb හික්මවනවා-veb
good(2-JJX) good-PRN
good(9-JJX) good-PRN
good(2-JJX) හොඳ-adj දක්ෂ-adj සොදුරු-adj කලාණ-adj යහපත්-adj හිතවත්-adj ඉටු-adj
තරමක-adj සුභ-adj ඉෂ්ට-adj
beautiful(5-JJX) ලස්සන-adj හැඩ-adj ලක්ෂණ-adj සොදුරු-adj කලාණ-adj අලංකාර-adj
රුමත්-adj විසිතුරු-adj සුමන-adj විචිත්‍ර-adj
read(7-JJX) කියවන ලද-adj
good(9-JJX) හොඳ-adj දක්ෂ-adj සොදුරු-adj කලාණ-adj යහපත්-adj හිතවත්-adj ඉටු-adj
තරමක-adj සුභ-adj ඉෂ්ට-adj
good(2-NWS) යහපත-nun ප්‍රයෝජනය-nun
boy(3-NWS) පිරිමි ළමයා-nun කොල්ලා-nun ළමයා-nun කුමාරයා-nun කොළුවා-nun බාලයා-
nun ගැටයා-nun වැඩ කරුවා-nun මානවකයා-nun ආවතේව කාරයා-nun
girl(6-NWS) ගැනු ළමයා-nun කෙල්ල-nun දැරිය-nun කුමරිය-nun යුවතිය-nun කුමාරිය-nun
පැටිකි-nun බාලිකාව-nun මානවිකාව-nun ළදැරිය-nun
good(9-NWS) යහපත-nun ප්‍රයෝජනය-nun
book(10-NWS) පොත-nun කෘතිය-nun පුස්තක-nun ග්‍රන්ථය-nun ප්‍රකරණය-nun
school(13-NWS) විද්‍යාස්ථානය-nun පාසැල-nun විද්‍යාලය-nun පාඨශාලාව-nun
විශ්වවිද්‍යාලයේ විද්‍යාංශය-nun ශික්ෂාලය-nun
the(1-DET) -det
good(2-RBX) සෘදු-adv
and(4-CON) සහ-con හා-con ද-con පිණිස-con ත්-con එවිට-con
a(8-DET) a-PRN
good(9-RBX) සෘදු-adv
at(11-PRP) දී-prp
the(12-DET) -det
read(7-VBP) කියවනවා-veb හදාරනවා-veb දක්වනවා-veb ඉගෙනගන්නවා-veb
read(7-VBD) කියවනවා-veb හදාරනවා-veb දක්වනවා-veb ඉගෙනගන්නවා-veb
read(7-VBN) කියවනවා-veb හදාරනවා-veb දක්වනවා-veb ඉගෙනගන්නවා-veb

++++
+ +
+ Phrase Generation +
+ +
++++

English phrase agent count 6

English Phrase Details

Counter : 0
EPH : 1: NP-(1, 3-DET,JJX,NWS) NOS, NWS
Start Pos 1

End Pos 3
MOP NWS

SWL?>පිරිමි ළමයා-nun කොල්ලා-nun ළමයා-nun කුමාරයා-nun කොළුවා-nun බාලයා-nun
ගැටයා-nun වැඩ කරුවා-nun මානවකයා-nun ආවතේව කාරයා-nun
TP NP

Sinhala Noun Phrase

1: NP-(1, 3-DET,JJX,NWS) NOS, NWS
HWL පිරිමි ළමයා-nun කොල්ලා-nun ළමයා-nun කුමාරයා-nun කොළුවා-nun බාලයා-nun
ගැටයා-nun වැඩ කරුවා-nun මානවකයා-nun ආවතේව කාරයා-nun

English Phrase Details

Counter : 1
EPH : 2: CN-(4, 4-CON) CON, CON
Start Pos 4
End Pos 4
MOP

SWL?>සහ-con හා-con ද-con පිණිස-con ත්-con එව්ව-con
TP CN

Sinhala Verb Phrase

2: CN-(4, 4-CON) CON, CON
HWL සහ-con හා-con ද-con පිණිස-con ත්-con එව්ව-con

English Phrase Details

Counter : 2
EPH : 3: NP-(5, 6-JJX,NWS) NOS, NWS
Start Pos 5
End Pos 6
MOP

SWL?>ගැනු ළමයා-nun කෙල්ල-nun දැරිය-nun කුමරිය-nun යුවතිය-nun කුමාරිය-nun
පැවික්කි-nun බාලිකාව-nun මානවිකාව-nun ළදැරිය-nun
TP NP

Sinhala Noun Phrase

3: NP-(5, 6-JJX,NWS) NOS, NWS
HWL ගැනු ලමයා-nun කෙල්ල-nun දැරිය-nun කුමරිය-nun යුවතිය-nun කුමාරිය-nun පැටිකි-
nun බාලිකාව-nun මානවිකාව-nun ළදැරිය-nun
+++++

English Phrase Details

Counter : 3
EPH : 4: VP-(7, 7-VBP) ACT-SPT, VBP
Start Pos 7
End Pos 7
MOP

SWL?>කියවනවා-veb හදාරනවා-veb දක්වනවා-veb ඉගෙනගන්නවා-veb
TP VP

+++++
Sinhala Verb Phrase
4: VP-(7, 7-VBP) ACT-SPT, VBP
HWL කියවනවා-veb හදාරනවා-veb දක්වනවා-veb ඉගෙනගන්නවා-veb
+++++

English Phrase Details

Counter : 4
EPH : 5: NP-(8, 10-DET,JJX,NWS) NOS, NWS
Start Pos 8
End Pos 10
MOP

SWL?>පොත-nun කෘතිය-nun පුස්තක-nun ග්‍රන්ථය-nun ප්‍රකරණය-nun
TP NP

+++++
Sinhala Noun Phrase

5: NP-(8, 10-DET,JJX,NWS) NOS, NWS
HWL පොත-nun කෘතිය-nun පුස්තක-nun ග්‍රන්ථය-nun ප්‍රකරණය-nun
+++++

English Phrase Details

Counter : 5
EPH : 6: PP-(11, 13-PRP,DET,NWS) NOS, NWS
Start Pos 11

End Pos 13

MOP

SWL?>විද්‍යාස්ථානය-nun පාසැල-nun විද්‍යාලය-nun පාඨශාලාව-nun විශ්වවිද්‍යාලයයේ
විද්‍යාංශය-nun ශික්ෂාලය-nun
TP PP

+++++

Sinhala PP Phrase

6: PP-(11, 13-PRP,DET,NWS) NOS, NWS

HWL විද්‍යාස්ථානය-nun පාසැල-nun විද්‍යාලය-nun පාඨශාලාව-nun විශ්වවිද්‍යාලයයේ විද්‍යාංශය-
nun ශික්ෂාලය-nun

+++++

Sinhala phrase agents count : 6

NP info : NWS3පිරිමි ළමයා-nun කොල්ලා-nun ළමයා-nun කුමාරයා-nun කොළුවා-nun
බාලයා-nun ගැටයා-nun වැඩ කරුවා-nun මානවකයා-nun ආවතේව කාරයා-nun

RRULE 101

MORP

RULE 101

OWORD: පිරිමි ළමයා

MORP CON-TS

Generated head word : පිරිමි ළමයා

OWORD: සහ

SMG (101) -CON-TS>සහ

RULE 226

MORP ACT-SPT-TS

OWORD: කියවනවා

SMG (226) -ACT-SPT-TS>කියවනවා

NP info : NWS6ගැනු ළමයා-nun කෙල්ල-nun දැරිය-nun කුමරිය-nun යුවතිය-nun කුමාරිය-
nun පැවික්කි-nun බාලිකාව-nun මානවිකාව-nun ළදැරිය-nun

RRULE 101

MORP

OWORD: ගැනු ළමයා

Generated head word : ගැනු ළමයා

RULE 101

MORP CON-TS

OWORD: හා

SMG (101) -CON-TS>හා

RULE 207

MORP ACT-SPT-TS

OWORD: හදාරනවා

SMG (207) -ACT-SPT-TS>හදාරනවා

RULE 101
 MORP CON-TS
 OWORD: ද
 SMG (101) -CON-TS>ද
 RULE 226
 MORP ACT-SPT-TS
 OWORD: දක්වනවා
 SMG (226) -ACT-SPT-TS>දක්වනවා
 RULE 101
 MORP CON-TS
 OWORD: පිණිස
 SMG (101) -CON-TS>පිණිස
 RULE 201
 MORP ACT-SPT-TS
 OWORD: ඉගෙනගන්නවා
 SMG (201) -ACT-SPT-TS>ඉගෙනගන්නවා
 RULE 101
 MORP CON-TS
 OWORD: ක්
 SMG (101) -CON-TS>ක්
 0.9284 : දක්වයි
 0.0597 : කියවයි
 0.0118 : හදාරයි
 0.0001 : ඉගෙනගයි
 #SP.4@VPA4: VP-(7, 7-VBP) ACT-SPT, VBP දක්වයි , කියවයි , හදාරයි , ඉගෙනගයි ,
 [SP.4@VPA] - - - update-sin-ph-action >> [genrate.701@sinph]
 SP.4@VPA I am the action verb read
 RULE 101
 MORP CON-TS
 OWORD: එව්ව

SP.4@VPAread I VPPPPPPPPPPPPPP ActionVerb
 SMG (101) -CON-TS>එව්ව
 0.4644 : සහ
 0.2533 : හා
 0.2256 : ද
 0.0201 : එව්ව
 0.0185 : පිණිස
 0.0182 : ක්
 [SP.2@CNA] - - - update-sin-ph-action >> [genrate.701@sinph]
 NP info : NWS10පොත-nun කෘතිය-nun පුස්තක-nun ග්‍රන්ථය-nun ප්‍රකරණය-nun

RRULE 102

MORP
OWORD: පොත
Generated head word : පොත
NP GEN101sdsv1විද්‍යාස්ථානය

RRULE 101
MORP NOS-TS
OWORD: විද්‍යාස්ථානය
Generated head word : විද්‍යාස්ථානය
NP GEN101adjදී
NP GEN101sdsv1පාසැල

RRULE 101
MORP NOS-TS
OWORD: පාසැල
Generated head word : පාසැල
NP GEN101adjදී
NP GEN101sdsv1විද්‍යාලය

RRULE 101
MORP NOS-TS
OWORD: විද්‍යාලය
Generated head word : විද්‍යාලය
NP GEN101adjදී
NP GEN101පිරිමි ළමයා

RRULE 101
MORP
OWORD: කොල්ලා
Generated head word : කොල්ලා
NP GEN101ගැනු ළමයා

RRULE 101
MORP
OWORD: කෙල්ල
Generated head word : කෙල්ල
NP GEN101sdsv1පාඨශාලාව

RRULE 101
MORP NOS-TS
OWORD: පාඨශාලාව
Generated head word : පාඨශාලාව
NP GEN101adjදී
NP GEN101sdsv1විශ්වවිද්‍යාලයයේ විද්‍යාංශය

RRULE 101
MORP NOS-TS
OWORD: විශ්වවිද්‍යාලයේ විද්‍යාංශය
Generated head word : විශ්වවිද්‍යාලයේ විද්‍යාංශය
NP GEN101adjදී
NP GEN102පොත

RRULE 101
MORP
OWORD: කෘතිය
Generated head word : කෘතිය
NP GEN101sdsv1ශික්ෂාලය

RRULE 101
MORP NOS-TS
OWORD: ශික්ෂාලය
Generated head word : ශික්ෂාලය
NP GEN101adjදී
0.8662 : විද්‍යාලයේ දී
0.1281 : පාසැලේ දී
0.0043 : පාඨශාලාවේ දී
0.0014 : විද්‍යාස්ථානයේ දී
0 : විශ්වවිද්‍යාලයේ විද්‍යාංශයේ දී
0 : ශික්ෂාලයේ දී

SP.6@PPA6: PP-(11, 13-PRP,DET,NWS) NOS, NWS විද්‍යාලය දී, පාසැල දී, පාඨශාලාව
දී, විද්‍යාස්ථානය දී, විශ්වවිද්‍යාලයේ විද්‍යාංශය දී, ශික්ෂාලය දී,
[SP.6@PPA] - - - update-sin-ph-action >> [generate.701@sinph]
[generate.701@sinph] [MP-OK]
[generate.701@sinph] ->(Sinhala phrase Manager) GET:[Message() from SP.2@CNA
- to generate.701@sinph - message - update-sin-ph-action - replyTo SP.2@CNA -
Ontology ema - content none - header manager - conid 0001 - data for info - ln EN-
16].
NP GEN101කොල්ලා

RRULE 101
MORP
OWORD: ළමයා
Generated head word : ළමයා
NP GEN101sdsv1කෙල්ල

RRULE 101
MORP
OWORD: දැරිය
Generated head word : දැරිය

[generate.701@sinph] ->(Sinhala phrase Manager) GET:[Message() from SP.4@VPA - to generate.701@sinph - message - update-sin-ph-action - replyTo SP.4@VPA - Ontology ema - content none - header manager - conid 0001 - data for info - In EN-16].

[generate.701@sinph] ->(Sinhala phrase Manager) GET:[Message() from SP.6@PPA - to generate.701@sinph - message - update-sin-ph-action - replyTo SP.6@PPA - Ontology ema - content none - header manager - conid 0001 - data for info - In EN-16].

NP GEN101කෘතිය

RRULE 101

MORP

OWORD: පුස්තක

Generated head word : පුස්තක

NP GEN101sdsv1ළමයා

RRULE 101

MORP

OWORD: කුමාරයා

Generated head word : කුමාරයා

NP GEN101sdsv1දැරිය

RRULE 101

MORP

OWORD: කුමරිය

Generated head word : කුමරිය

NP GEN101sisv1පුස්තක

RRULE 101

MORP

OWORD: ගුණ්ථය

Generated head word : ගුණ්ථය

NP GEN101sdsv1කුමරිය

RRULE 101

MORP

OWORD: යුවතිය

Generated head word : යුවතිය

NP GEN101sdsv1කුමාරයා

RRULE 101

MORP

OWORD: කොළුවා

Generated head word : කොළුවා

NP GEN101sisv1ගුණ්ථය

RRULE 101
MORP
OWORD: ප්‍රකරණය
Generated head word : ප්‍රකරණය
NP GEN101sdsv1සුවතිය

RRULE 101
MORP
OWORD: කුමාරිය
Generated head word : කුමාරිය
NP GEN101sdsv1කොළුවා

RRULE 101
MORP
OWORD: බාලයා
Generated head word : බාලයා
NP GEN101sisv1ප්‍රකරණය
0.8222 : හොඳ පොතක්
0.0718 : හොඳ ග්‍රන්ථයක්
0.0631 : හොඳ කෘතියක්
0.0397 : හොඳ පුස්තකක්
0.0032 : හොඳ ප්‍රකරණයක්

SP.5@NPA5: NP-(8, 10-DET,JJX,NWS) NOS, NWS හොඳ පොත , හොඳ ග්‍රන්ථයක් ,
හොඳ කෘතිය , හොඳ පුස්තක , හොඳ ප්‍රකරණයක් ,
[SP.5@NPA] - - - update-sin-ph-action >> [genrate.701@sinph]

SP.5@NPAa good book I AM NPnull
NP GEN101sdsv1කුමාරිය

RRULE 101
MORP
OWORD: පැටිකි
Generated head word : පැටිකි
NP GEN101sdsv1බාලයා

RRULE 101
MORP
OWORD: ගැටයා
Generated head word : ගැටයා
[genrate.701@sinph] -(Sinhala phrase Manager) GET:[Message() from SP.5@NPA
- to genrate.701@sinph - message - update-sin-ph-action - replyTo SP.5@NPA -

Ontology ema - content none - header manager - conid 0001 - data for info - In EN-16].

NP GEN101sds1ඉවසා

RRULE 101

MORP

OWORD: වැඩ කරුවා

Generated head word : වැඩ කරුවා

NP GEN101sds1පැවික්කි

RRULE 101

MORP

OWORD: බාලිකාව

Generated head word : බාලිකාව

NP GEN101sds1වැඩ කරුවා

RRULE 101

MORP

OWORD: මානවකයා

Generated head word : මානවකයා

NP GEN101sds1බාලිකාව

RRULE 101

MORP

OWORD: මානවිකාව

Generated head word : මානවිකාව

NP GEN101sds1මානවකයා

RRULE 101

MORP

OWORD: ආවණේව කාරයා

Generated head word : ආවණේව කාරයා

NP GEN101sds1මානවිකාව

RRULE 101

MORP

OWORD: ළදැරිය

Generated head word : ළදැරිය

NP GEN101sds1ආවණේව කාරයා

0.5302 : හොඳ ළමයා

0.3938 : හොඳ කොල්ලා

0.0311 : හොඳ කුමාරයා

0.0117 : සුභ කොලුවා

0.0098 : හොඳ වැඩ කරුවා

0.0095 : හොඳ බාලයා

0.0061 : හොඳ ගැටයා
 0.0058 : හොඳ පිරිමි ලමයා
 0.0017 : යහපත් මානවකයා
 0.0004 : හොඳ ආවණේව කාරයා

SP.1@NPA1: NP-(1, 3-DET,JJX,NWS) NOS, NWS හොඳ ළමයා , හොඳ කොල්ලා , හොඳ කුමාරයා , සුභ කොලුවා , හොඳ වැඩ කරුවා , හොඳ බාලයා , හොඳ ගැටයා , හොඳ පිරිමි ලමයා , යහපත් මානවකයා , හොඳ ආවණේව කාරයා ,
 [SP.1@NPA] - - - update-sin-ph-action >> [genrate.701@sinph]

SP.1@NPAthe good boy I AM NPnull
 NP GEN101sdsv1 ළදරිය

0.621 : ලස්සන දැරිය
 0.2074 : ලස්සන කෙල්ල
 0.1079 : ලස්සන කුමරිය
 0.0266 : ලස්සන යුවතිය
 0.0174 : ලස්සන ගැනු ළම
 0.0108 : ලස්සන පැටිකි
 0.0076 : ලස්සන කුමාරිය
 0.0006 : ලස්සන බාලිකාව
 0.0004 : ලස්සන ළදරිය
 0.0003 : ලස්සන මානවිකාව

SP.3@NPA3: NP-(5, 6-JJX,NWS) NOS, NWS ලස්සන දැරිය , ලස්සන කෙල්ල , ලස්සන කුමරිය , ලස්සන යුවතිය , ලස්සන ගැනු ළම , ලස්සන පැටිකි , ලස්සන කුමාරිය , ලස්සන බාලිකාව , ලස්සන ළදරිය , ලස්සන මානවිකාව ,
 [SP.3@NPA] - - - update-sin-ph-action >> [genrate.701@sinph]

SP.3@NPAbeautiful girl I AM NPnull

[genrate.701@sinph] ->(Sinhala phrase Manager) GET:[Message() from SP.1@NPA - to genrate.701@sinph - message - update-sin-ph-action - replyTo SP.1@NPA - Ontology ema - content none - header manager - conid 0001 - data for info - ln EN-16].

[genrate.701@sinph] ->(Sinhala phrase Manager) GET:[Message() from SP.3@NPA - to genrate.701@sinph - message - update-sin-ph-action - replyTo SP.3@NPA - Ontology ema - content none - header manager - conid 0001 - data for info - ln EN-16].

EnSiMaS Phrase list

- 1: NP-(1, 3-DET, JJX, NWS) NOS, NWS හොඳ ළමයා , හොඳ කොල්ලා , හොඳ කුමාරයා , සුභ කොළුවා , හොඳ වැඩ කරුවා , හොඳ බාලයා , හොඳ ගැටයා , හොඳ පිරිමි ළමයා , යහපත් මානවකයා , හොඳ ආවර්ණික කාරයා ,
- 2: CN-(4, 4-CON) CON, CON සහ , හා , ද , එවිට , පිණිස , ත් ,
- 3: NP-(5, 6-JJX, NWS) NOS, NWS ලස්සන දැරිය , ලස්සන කෙල්ල , ලස්සන කුමරිය , ලස්සන යුවතිය , ලස්සන ගැනු ළමයා , ලස්සන පැටිකි , ලස්සන කුමාරිය , ලස්සන බාලිකාව , ලස්සන ළදැරිය , ලස්සන මානවිකාව ,
- 4: VP-(7, 7-VBP) ACT-SPT, VBP දක්වයි , කියවයි , හදාරයි , ඉගෙනගනියි ,
- 5: NP-(8, 10-DET, JJX, NWS) NOS, NWS හොඳ පොත , හොඳ ග්‍රන්ථයක් , හොඳ කෘතිය , හොඳ පුස්තකය , හොඳ ප්‍රකරණයක් ,
- 6: PP-(11, 13-PRP, DET, NWS) NOS, NWS විද්‍යාලයේදී , පාසැලේ දී , පාඨශාලාවේ දී , විශ්වවිද්‍යාලයේ විද්‍යාංශය දී , ශික්ෂාලයේ දී ,
- Verb Position 3
Subject Position 2

#දැරිය දක්වනවා	0.562
#දැරිය කියවනවා	0.366
#දැරිය හදාරනවා	0.043
#දැරිය ඉගෙනගන්නවා	0.028

#කෙල්ල කියවනවා	0.726
#කෙල්ල දක්වනවා	0.17
#කෙල්ල හදාරනවා	0.056
#කෙල්ල ඉගෙනගන්නවා	0.048

#කුමරිය දක්වනවා	0.41
#කුමරිය කියවනවා	0.362
#කුමරිය හදාරනවා	0.169
#කුමරිය ඉගෙනගන්නවා	0.059

#යුවතිය කියවනවා	0.518
#යුවතිය දක්වනවා	0.39
#යුවතිය ඉගෙනගන්නවා	0.062
#යුවතිය හදාරනවා	0.03

#ගැනු ළමයා කියවනවා	0.626
#ගැනු ළමයා දක්වනවා	0.251
#ගැනු ළමයා හදාරනවා	0.073
#ගැනු ළමයා ඉගෙනගන්නවා	0.05

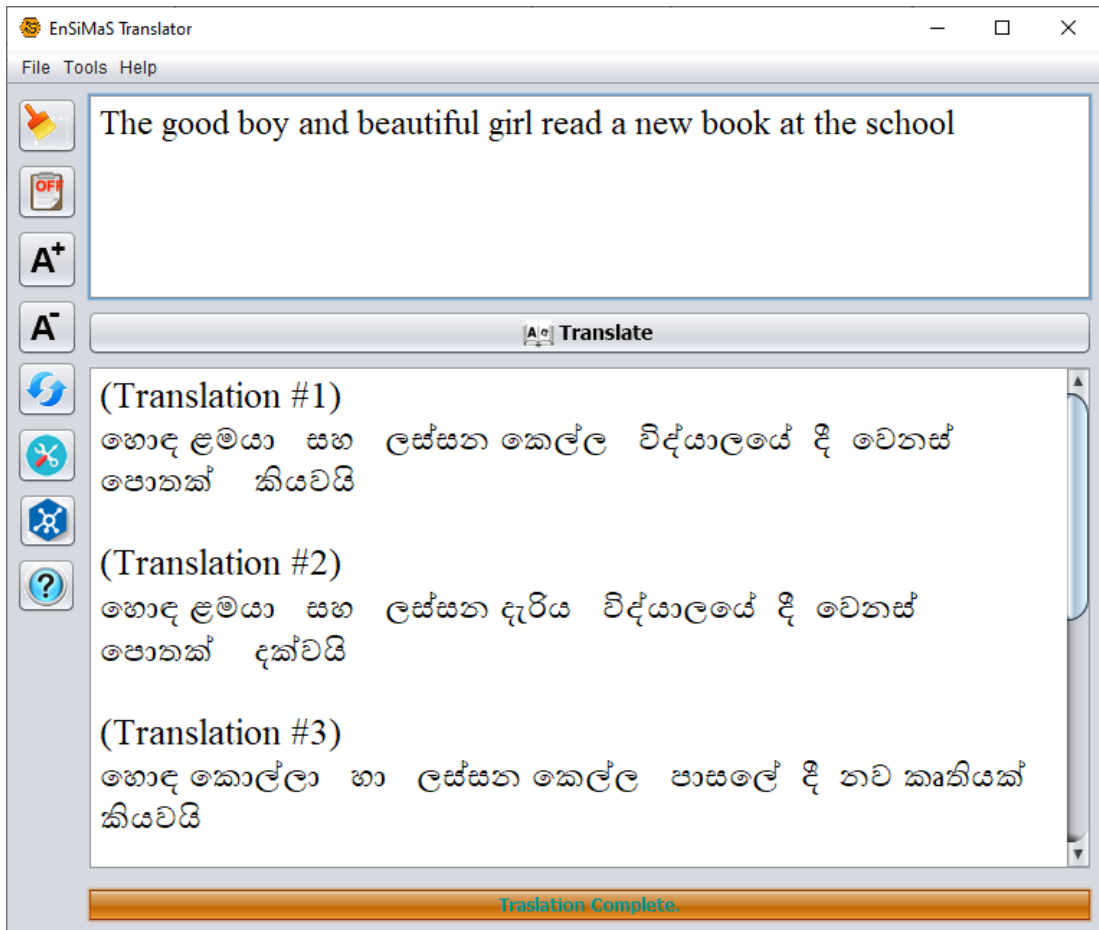
#පැටික්කි කියවනවා	0.694
#පැටික්කි ඉගෙනගන්නවා	0.18
#පැටික්කි දක්වනවා	0.126
#පැටික්කි හදාරනවා	0

#කුමාරිය දක්වනවා	0.488
#කුමාරිය කියවනවා	0.377
#කුමාරිය හදාරනවා	0.093
#කුමාරිය ඉගෙනගන්නවා	0.041

#බාලිකාව දක්වනවා	0.423
#බාලිකාව කියවනවා	0.311
#බාලිකාව හදාරනවා	0.265
#බාලිකාව ඉගෙනගන්නවා	0

#ලදැරිය දක්වනවා	0.882
#ලදැරිය කියවනවා	0.097
#ලදැරිය හදාරනවා	0.022
#ලදැරිය ඉගෙනගන්නවා	0

#මානවිකාව කියවනවා	0.935
#මානවිකාව දක්වනවා	0.065
#මානවිකාව හදාරනවා	0
#මානවිකාව ඉගෙනගන්නවා	0



Appendix B: EnSiMaS User Manual

System Overview

EnSiMaS is fully Java-based English to Sinhala Machine Translation system. EnSiMaS also a multi-agent system should capable to communicate with each other and provides accepted translations for given English text. The system uses Multi agent and phrase-based psycholinguistics parsing approach to provide appropriate Sinhala translations.

System features

In general, the system consists of the following features

1. The system is fully Java-based Therefore it should capable to run on any Machine
2. Provide Grammatically correct translation with Sinhala Morphological generation
3. Provide Multiple solutions
4. Can customize the Ontology for better results

System Requirements

The following is the software and hardware specification for the EnSiMaS

- Any Operating System with JRE 1.7 or above and Sinhala Unicode support
- Internet connection for a better solution
- 2GB Ram or above

Installation

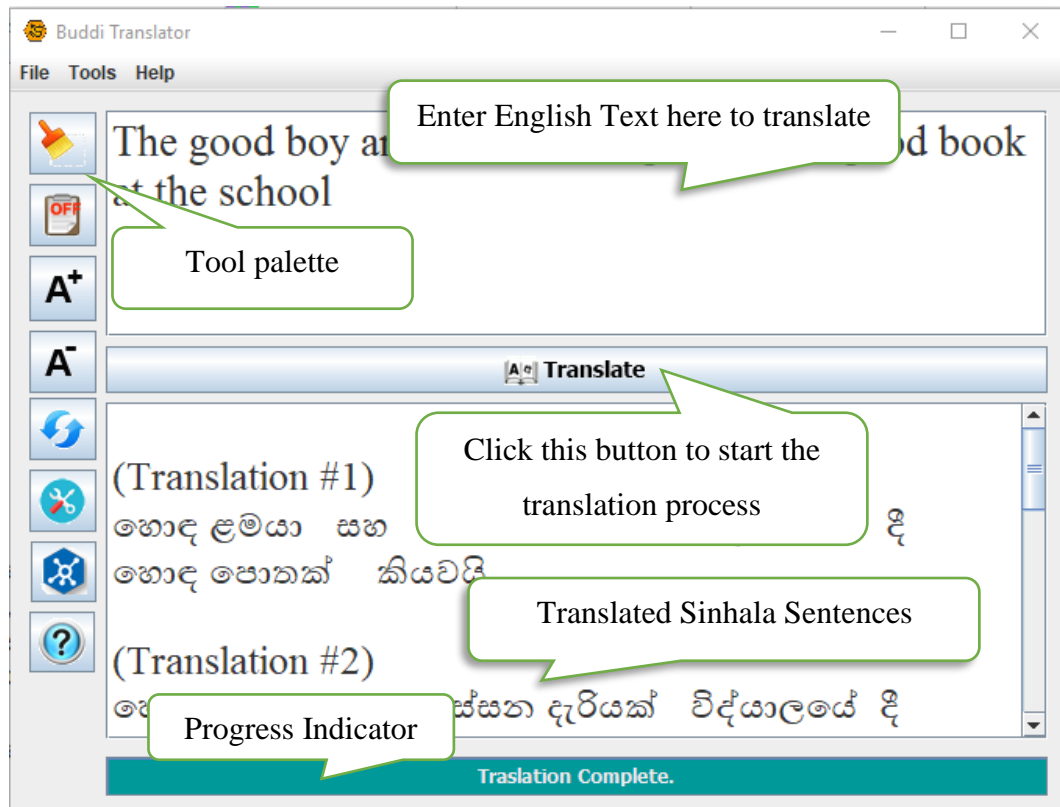
No need to install just click on the EnSiMaS.jar (or user can execute RUN.bat). then the system should appear the Main translation interface.

User Interfaces

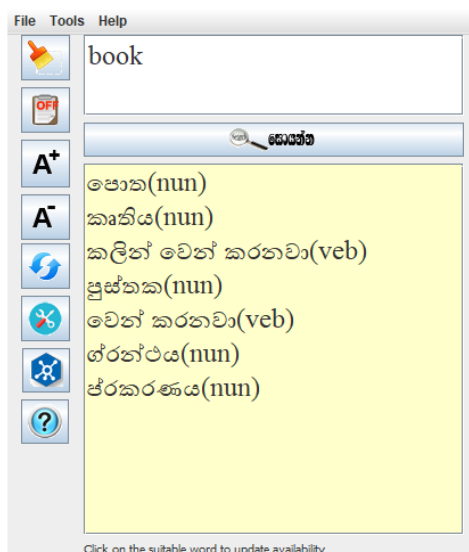
Translator consists of a simple GUI with some useful supporting features. Main translation system should consist of Main menu, ToolPanel, Input and output text fields.

EnSiMaS Translator (Buddi Translator)

The following figure shows the user interface of the EnSiMAS.

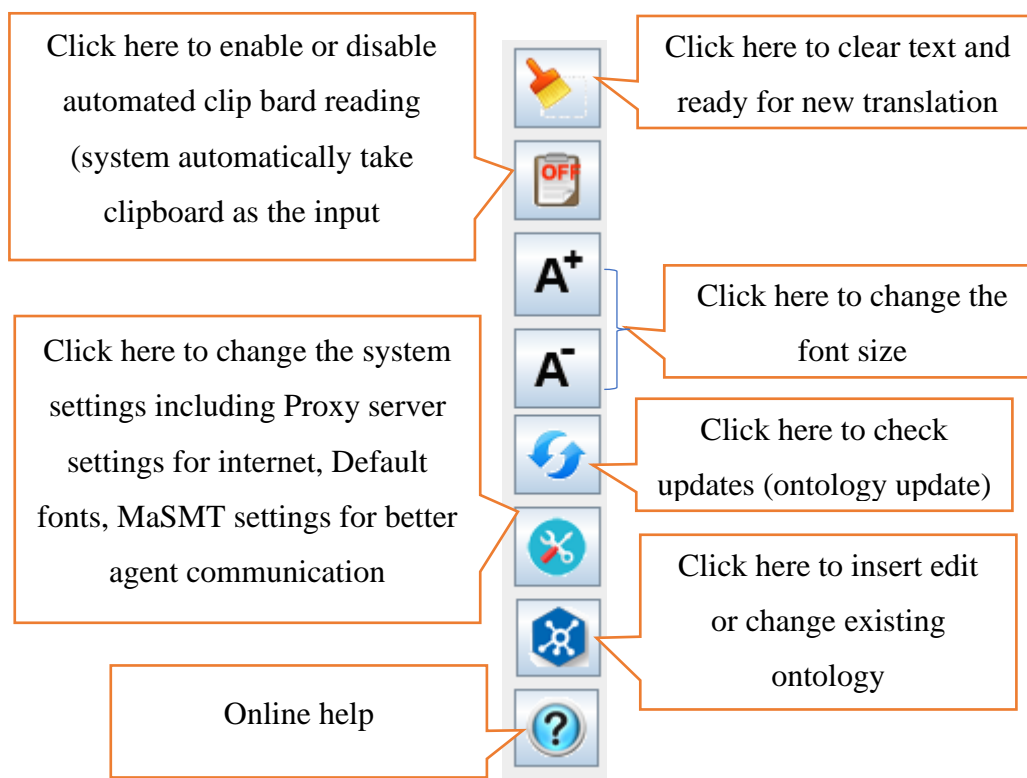


EnSiMaS Dictionary



Tool Platte

The tool palette consists of 8 components to provide an easy way to translation.



Appendix C: Sample of evaluation form

3

Form for Evaluate English to Sinhala Translations

1. This evaluation form is to evaluate the English to Sinhala translation done through the Machine Translation system.

2. We have provided 3 translations for 25 English Sentences which was translated by the system and a human expert.

3. In here we request to take your response (Adequacy and Fluency) for these three Sinhala translations for the given English sentence.

4. Please enter Adequacy and Fluency values (1-5 scale) on provided space.

Adequacy (Source to Target Meaning translation)

- 1- None (කිසිම තේරුමක් නැත)
- 2- Little Meaning (ප්‍රමාණවත් අදහසක් නැත)
- 3- Much Meaning (සම් අදහසක් ලැබේ)
- 4- Most Meaning (තේරුම් ගතහැකිය)
- 5- All Meaning (තේරුම ඉතාම හොඳින් ඇත)

Fluency (සිංහල පරිවර්තනයේ නිවැරදි බව)

- 1- Incomprehensible (පමිදුරුකොන් වැරදි)
- 2- Diffluent Sinhala (වැරදි කිහිපයක් ඇත)
- 3- Non-native Sinhala (සාමාන්‍යයි)
- 4- Good Sinhala (හොඳයි)
- 5- Flawless Sinhala (ඉතා නිවැරදි)

NO	English Sentence	Translation 1		Translation 2		Translation 3		Adequacy	Fluency
		Adequacy	Fluency	Adequacy	Fluency	Adequacy	Fluency		
Example									
	The boy reads a book	5	5	5	5	5	5	4	4
Translations									
1	I am writing a book	5	5	5	4	5	5	5	5
2	The good boy reads a new book	3	3	3	5	5	5	5	4
3	A good student will eat rice at the canteen	3	4	3	3	3	5	5	5
4	The man with his wife went to the party	5	4	5	5	5	5	4	3
5	The good boy was eating an apple at the school	5	5	5	5	5	5	4	4
6	The good boy and a beautiful girl have written an essay	5	5	5	5	5	5	4	3

7	I will write a story for my children	මම මගේ ළමයින් සඳහා කතාවක් ලියන්නෙමි	4	4	මම මගේ දරුවන්ට කතාවක් ලියන්නෙමි	5	4	මම මගේ ළමයි සඳහා කතාවක් රචනා කරන්නෙමි	4	9
8	I play basketball every week with my friends	මම මගේ මිතුරන් සමඟ සැම සතියකම පැයින්ද ක්‍රීඩා කරමි	4	4	මම සෑම සතියකම මගේ මිතුරන් සමඟ පැයින්ද ක්‍රීඩා කරමි	5	4	මම මගේ මිතුර සමඟ සෑම සතියකම පැයින්ද ක්‍රීඩා කරමි	5	9
9	My good friend was singing a song at the school	මගේ හොඳ මිතුර පාසලේ ගීතයක් ගායනා කරමින් සිටියේය	4	3	මගේ හොඳ මිතුර පාසලේ ගීතයක් ගායනා කරමින් සිටියේය	4	3	මගේ හොඳ මිතුර විද්‍යාලයේ දී ගීතයක් ගායනා කරමින් සිටියේය	5	4
10	A boy and a girl sing a song on the bus	පිරිමි ළමයෙක් සහ ගැහැණු ළමයෙක් බස් රථයේ ගීතයක් ගායනා කරති	4	4	පිරිමි ළමයෙක් සහ ගැහැණු ළමයෙක් බස් රථයේ ගීතයක් ගායනා කරති	5	4	පිරිමි ළමයෙක් සහ ගැහැණු ළමයෙක් බස් රථයේ ගීතයක් ගායනා කරති	5	4
11	the boy is going to eat rice with my friend	පිරිමි ළමයා මගේ මිතුර සමඟ බත් කන්න යන්නේය	4	3	පිරිමි ළමයා මගේ මිතුර සමඟ බත් කන්න යනවා	4	3	කොල්ලා මගේ මිතුර සමඟ බත් අනුභව කරන්නේය	3	9
12	I wrote a letter at the school	මම පාසලේදී ලිපියක් ලිවුවෙමි	4	3	මම පාසලේදී ලිපියක් ලිවුවා	5	3	මම පාසලේදී ලිපියක් රචනා කළෙමි	4	9
13	We read newspapers daily	අපි දිනපතා පුවත්පත් කියවමු	3	3	අපි දිනපතා පුවත්පත් කියවනවා	5	3	අපි ප්‍රචානි පත්‍ර දිනපතා කියවමු	4	3
14	She can ride a bicycle and drive a car	ඇයට බයිසිකලයක් පැදවෙනවා සහ වාහනයක් පදවන්න පුළුවන්	4	4	ඇය පදවෙනවා බයිසිකලයක් පැදවෙනවා සහ වාහනයක් පදවන්න පුළුවන්	1	1	ඇයට බයිසිකලයක් පැදවෙනවා සහ වාහනයක් පදවන්න පුළුවන්	3	9
15	Mother prepares the breakfast at the kitchen for her children	මව තම දරුවන් සඳහා මුළුතැන්ගෙයෙහි උදෑසන ආහාරය පිළියෙළ කරන්නීය	5	5	මව තම දරුවන් සඳහා මුළුතැන්ගෙයෙහි උදෑසන ආහාරය පිළියෙළ කරයි	5	5	මව කුස්සියේ දී ඇයගේ ළමයි සඳහා උදෑසන ආහාරය පිළියෙළ කරයි	3	3
16	My dog ate rice with meat	මගේ බල්ලා මංච සමඟ බත් කෑවේය	5	4	මගේ බල්ලා මත් සමඟ බත් කෑවා	4	4	මගේ බල්ලා මංච සමඟ බත් කෑවේය	5	4
17	I was eating a big pizza with my friends	මම මගේ මිතුරන් සමඟ පොකු පිසා ඊකක් කමින් සිටියෙමි	5	5	මම මගේ මිතුරන් සමඟ පොකු පිසා කමින් සිටියෙමි	3	3	මම මගේ මිතුරන් සමඟ විදුලන් ඉතාලියානු ආහාර විශේෂයක් කමින් සිටියෙමි	3	9

18	I am selling my motorcycle and buying a new car	මම මගේ බයිසිකලය විකුණා අලුත් කාර් එකක් මිලදී ගනිමි	3	3	මම මගේ බයිසිකලය විකුණා අලුත් කාර් එකක් මිලදී ගන්නවා	2	14	මම මගේ බයිසිකල විකුණනවා සහ නව රථයක් මිලට ගන්නවා	5	5
19	We wrote a book	අපි පොතක් ලිව්වෙමු	5	5	අපි පොතක් ලිව්වා	5	5	අපි පොතක් රචනා කළෙමු	5	5
20	My dog and his cat are eating rice with meat	මගේ බල්ලා සහ ඔහුගේ බලලා මස් සමඟ බිත් කමින් සිටියි	3	3	මගේ බල්ලා සහ ඔහුගේ බලලා මස් සමඟ බිත් කනවා	3	5	මගේ බල්ලා සහ ඔහුගේ පුංචා මාංශ සහ බිත් කමින් සිටියි	5	5
21	A clever student reads good newspapers daily	දක්ෂ ශිෂ්‍යයෙක් දිනපතා හොඳ පුවත්පත් කියවයි	5	5	දක්ෂ ශිෂ්‍යයෙක් දිනපතා හොඳ පුවත්පත් කියවනවා	4	4	දක්ෂ ශිෂ්‍යයෙක් හොඳ පුවත්පත් දිනපතා කියවයි	5	5
22	The singer has sung a new song	ගායකයා නව ගීතයක් ගායනා කර තිබේ	5	5	ගායකයා නව ගීතයක් ගායනා කර තිබේ	5	5	ගායකයා නව ගීතයක් ගායනා කරලා තියෙයි	4	5
23	The neighbour bought a radio	අසල්වැසියා ඉවන්විදුලියක් ගන්නෙයි	5	5	අසල්වැසියා ඉවන්විදුලියක් මිලදී ගන්නෙයි	5	5	අසල්වැසියා රේඩියෝවක් මිලට ගන්නෙයි	5	5
24	we will draw a painting at every weekends with our friends	අපි සෑම සති අන්තයකම අපගේ මිතුරන් සමඟ චිත්‍රයක් අඳින්නෙමු	2	2	අපි සෑම සති අන්තයකම අපගේ මිතුරන් සමඟ චිත්‍රයක් අඳින්නෙමු	3	3	අපි සෑම සති අන්තයකදීම සමඟ අපගේ මිතුරන් සමඟ සිත්තමක් අඳිමු	3	4
25	The strongest rain ever recorded in India shut down the financial hub of Mumbai, snapped communication lines, closed airports and forced thousands of people to sleep in their offices or walk home during the night, officials said today	ඉන්දියාවේ මෙතෙක් වාර්තාවූ දැඩි වර්ෂාව හේතුවෙන් මුම්බායි හි මූල්‍ය මධ්‍යස්ථානය වසා දැමීමත්, සන්නිවේදන මාර්ග අනාභිවිමත්, ඉවත් කොටුවලට වසා දැමීමත්, දහස් ගණනින් ජනයා කාර්යාලවල කොටුම්භවත් හෝ රාත්තරියේදී නිවෙස් බලා සිටත් විමටත් සිදු වූ බව අද බලධාරීන් පැවසූහ	4	5	ඉන්දියාවේ මෙතෙක් වාර්තාවූ මුම්බායි හි මූල්‍ය මධ්‍යස්ථානය වසා දැමීම, සන්නිවේදන මාර්ග, ඉවත් කොටුවලට වසා දැමීම සහ ජනතාවට නම් කාර්යාලවල නිදා ගැනීමත් හෝ රාත්තරියේදී නිවෙස් බලා සිටත් විමටත් සිදු වූ බව අද බලධාරීන් පැවසූහ	5	5	India හිදී ලියාපදිංචි කරන ලද දැඩි වර්ෂාව Mumbai හි මුදල් සිලිබරු මධ්‍යස්ථානය වසා සන්නිවේදන මාර්ග ආවරණය කළේය. ඉවත්කොටුපස සහ ඔවුන්ගේ කාර්යාලය හිදී ජනතාව නින්දාට හෝ රාත්‍රිය අතරතුර ගමන් කර නිලධාරීන් අද ප්‍රකාශ කළේය	2	2

Appendix D: List of Publications

1. B. Hettige, A. S. Karunananda, G. Rzevski, Multi-agent solution for managing complexity in English to Sinhala Machine Translation, *International Journal of Design & Nature and Ecodynamics*, Volume 11, Issue 2, 2016, 88 – 96.
2. B. Hettige, A. S. Karunananda, G. Rzevski, ” MaSMT: A Multi-agent System Development Framework for English-Sinhala Machine Translation”, *International Journal of Computational Linguistics and Natural Language Processing (IJCLNLP)*, Volume 2 Issue 7 July 2013.
3. B. Hettige, A. S. Karunananda, G. Rzevski, MaSMT4: The AGR Organizational Model-Based Multi-agent System Development Framework for Machine Translation, Accepted to present, the third International Conference, SLAAI-ICAI 2019, Sri Lanka, December 12, 2019.
4. B. Hettige, A. S. Karunananda, G. Rzevski, Thinking Like Humans: A New Approach to Machine Translation, *Proceedings of the Second International Conference, SLAAI-ICAI 2018*, Moratuwa, Sri Lanka, December 20, 2018, Springer Singapore
5. B. Hettige, A. S. Karunananda, G. Rzevski, Phrase-level English to Sinhala Machine Translation with Multi-Agent Approach, *Proceedings of the IEEE International Conference on Industrial and Information Systems (ICIIS 2017)*, Sri Lanka, 2017.
6. B. Hettige, A. S. Karunananda, G. Rzevski, A Sinhala Ontology Generator for English to Sinhala Machine Translation, *Proceedings of KDU International Research Symposium 2014*
7. B. Hettige, A. S. Karunananda, G. Rzevski, “Multi-agent System Technology for Morphological Analysis”, *Proceedings of the 9th Annual Sessions of Sri Lanka Association for Artificial Intelligence (SLAAI)*, Colombo, 2012.

Appendix E:
MaSMT Development Guide