Varanageema: A Theoretical basics for English to Sinhala Machine Translation

B. Hettige[1], A. S. Karunananda[2]
[1]Department of Statistics and Computer Science, Faculty of Applied Science,
University of Sri Jayewardenepura, Sri Lanka.
[2]Faculty of Information Technology, University of Moratuwa, Sri Lanka.
budditha@yahoo.com[1], asoka@itfac.mrt.ac.lk[2]

***Abstract* – Grammatically correct target language word generation is a research challenging task in any machine translation. To achieve this challenge, most of the machine translation systems were used theoretical basics with rule-based approaches. This paper reports the theoretical basics of the English to Sinhala machine translation named as concept of Varanegeema. The English to Sinhala machine translation system has been developed as a multi agent system and it provides grammatically correct Sinhala translation for given English text. This system uses the concept of Varanegeema in Sinhala language as theoretical basics for English to Sinhala machine translation. This Varanegeema concept also reduces the Sinhala word availability in the Sinhala dictionary and gives the way to generate grammatically correct Sinhala words. The translation system has been successfully tested through the limited domain and the obtained results that we consider very satisfactory. Details of the concept of Varanegeema, design, implementation and evaluation of the English to Sinhala machine translation are also given in the paper.**

I INTRODUCTION

The Machine Translation system is a type of computer software which is used to translate text or voice from one natural language into another [14]. Large number of Machine Translation systems have been developed, using several approaches including Human-assisted[13], Rule-based, Corpus-based, Knowledge-based[16][17], Hybrid, Agent-based [24][5] etc. Most of these machine translation systems were used theoretical basics with rule-based approaches to generate grammatically correct translation.

The rule based approach gives grammatical correct translation by using set of rules. Basically, the rule-based machine translation system contains a source language morphological analyzer, a source language parser, translator, target language morphological generator, target language parser and several lexicon dictionaries. Source language morphological analyzer analyzes a source language word and provides the morphological information. Source language parser is a syntax analyzer that analyzes source language sentences. Translator is used to translate a source language word into target language. Target language morphological generator generates appropriate target language words for the given grammatical information.

Also target language parser works as a composer and it composes a suitable target language sentence. Furthermore, this type of machine translation system needs minimum of three dictionaries namely, the source language dictionary, the bilingual dictionary and the target language dictionary. The source language morphological analyzer needs a source language dictionary for morphological analysis. Bilingual dictionary is used by the translator for translating source language into target language; and the target language morphological generator uses the target language dictionary to generate target language words. Most of the rule based machine translation systems were used theoretical basics for target language morphological generators. Figure 1 shows overview of the general rule-based machine translation system.

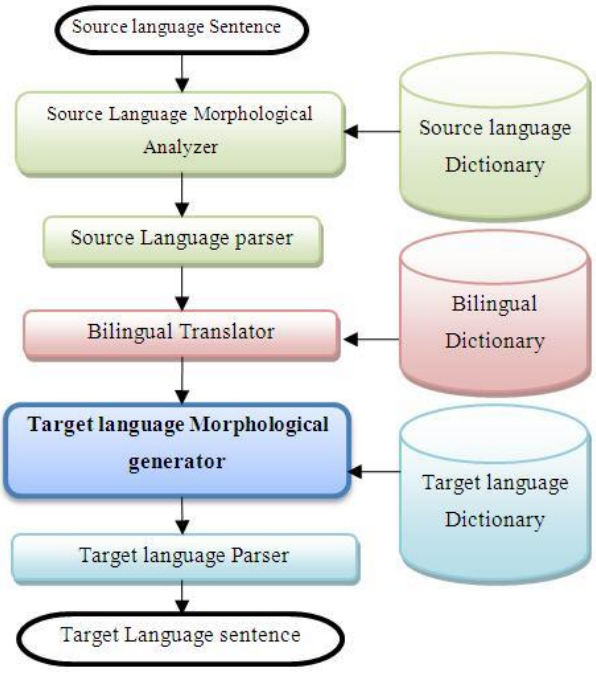

Figure 1: Overview of the general rule based machine translation system.

In the region, numbers of machine translation systems have been developed through the rule-based approach

with the theoretical basics. Many of these machine translation systems[2][4] are used Panini grammar [1] as the theoretical basics of the machine translation. For example the Anusararaka [2] Machine translation system has been used several Morphological generators for Indian family of machine translation. All of these language generators are designed through the Panini grammar framework [3][23].

Sinhala is one of Indo Ariean family language and it is the spoken language of 74% of the people in Sri Lanka. Sinhala is one of the constitutionally-recognized official languages of Sri Lanka, along with Tamil language. Sinhala has its own writing system, which is an offspring of the Brahmi script [11]. Maldives, Dhivehi are the closest relative languages for Sinhala. However, Sinhala differs from all other Indo-Aryan languages. In that, it contains a pair of vowel sounds that are unique to it, such as short vowel: 'ඇ' – ae and Long vowel: 'ඈ' – aae [21]. Also Sinhala contains a set of five nasal sounds known as "half nasal" or "prenasalized stops". These sounds as represented in modern Sinhala writing and their Romanized notation are as follows: ඟ (nng), `ඦ (ndj), `ඬ (nnd), ඳ (nd) and ඹ(mb) [20][22].

According to the Sinhala language structures, other language generators such as Hindi, English etc., directly cannot be used for Sinhala word generation. Therefore we have designed and developed a new word generator for Sinhala language named as Sinhala morphological generator. This Sinhala Morphological generator uses the concept of Varanegeema in Sinhala language as the theoretical basics of the word generation. By using this Sinhala morphological generator we have design and develop Multi agent based English to Sinhala machine translation system which translates English Sentence into Sinhala.

This paper presents the theoretical basics of the English to Sinhala machine translation named as the concept of "Varanageema" including design, implementation of the English to Sinhala machine translation system.

The rest of this paper is organized as follows. Section 2 describes overview of the Concept of Varanageema. Then Section 3 describes overview of English to Sinhala machine translation system. Session 4 describes more details about Sinhala word generation; Section 5 reports the evaluation methodology with the experimental results. Finally section 6 concludes the paper with conclusion and a note on further work.

II. THE CONCEPTS OF VARANAGEEMA

The concepts of Varanageema in the Sinhala language is the modification of the nouns, verbs form their based form. Basically a Sinhala noun contains more than 27 conjugation forms and more than 36 conjugation forms for a Sinhala verb[19]. Brief description of the Sinhala Nouns and the Sinhala Verbs are given below.

The Sinhala Noun Morphology

The Sinhala Noun is a word that represents the noun, pronoun and the adjective in the English language. The Sinhala noun has four types of inflections such as Gender (lingaya), Number (Wachana), Person (Purusha) and Case (Vibhakthi). There are three genders namely masculine gender, feminine gender and neuter gender. Singular and plural are the Number and there are three persons namely 1st person (Uththama purusha) 2nd person (Maddyama purusha) and 3rd person (prathama purusha). Also there are nine cases in Sinhala such as Nominative (prathama), Accusative (karma), Instrumental (kaththr), Auxiliary (karana), Dative (sampadana), Abilitive (avadi), Genitive (Sambanda), Locative (adara) and Vocative (alapana).[18][22] There are 27 inflection forms generated for single base noun such as nine Vibahakthi, article and the number. For example Sinhala base word '.j' inflects as '.jhd, .jfhda, .jfhla etc. The base word is directly affected by the nine cases. Some case suffixes are written with the base word and some are written separately. Table 4 shows sample case makers of the Sinhala noun. There are number of case makers forms available in Sinhala that depends on the gender of the noun.

From morphological point of view, a Sinhala noun is also a word, and nouns are participated inflection and derivations. The Sinhala nouns can be deivided into three categories, namely, simple, complex and compound. A Simple noun contains only a prakurthi (base form) while a complex noun contains prakruthi and suffix [21]. A compound noun contains two or more prakruthi. Prakurthi is a base form of a word and it is also in the non-inflection form.

As mentioned, Sinhala nouns can be devided into three categories. A brief description of these categories is as follows. A simple noun contains only nama prakurthi[20]. Nama prakurthi is a prakurthi, out of five prakurthis, such as, nama, kriya, guna vilasa and nipatha. Nama prakurthi is an adjective form of a noun. Nama prattya, vibakthi pratthaya, upasarga and thadhitha are used to inflect prakurthi. Upasarga is a prefix and others are suffixes. The following example shows how nouns are participated in inflection.

මිනිසා = මිනිස් + ආ = Prakurthi + Nama prathya
මිනිසාට = මිනිස් + ආ + ට = Prakurthi + Nama prathya + Vibakthi prattya
නොමිනිස් = නො +මිනිස් = Upasarga + Prakurthi

Note that, in the above the word මිනිස් is a prakurthi and 'ආ' is a nama prathya ,'ට' is a vibakthi suffix, 'නො' is a upasarga. The nama prakurtiya is a base form and nama prathya is one of the inflection parts of the noun. Also vibakthi suffix is an inflection part. There are 27 forms of noun that can be generated by inflecting a single root word (prakurthi). This inflecting

is called 'Nama varanagilla'(Word conjugation). The Sinhala noun contains more than hundred rules to conjugate a noun using a given base form (Prakurthi). In Sinhala there are 15 conjugation patterns identified for generating a Sinhala noun. These patterns are called 'Gana'. There are six noun generation forms (aeth ganaya, ali ganaya, tara ganaya, vasu ganaya, kaputu ganaya and bamara ganaya) that used to generate masculine gender nouns. There are nine generation forms (poth, akshara, basha, pili, akuru, polo, sulan, nuwara and mutu) that used to generate neuter gender[22]. The table 1 shows the conjugation table of the Sinhala word මුව ගණය

විභක්තිය Case	නියත ඒක Direct	අනියත ඒක Indirect	බහු වචන Plural
ප්‍රථමා	මුවා	මුවෙක්	මුවෝ
කර්ම	මුවා	මුවකු	මුවෝ
කතෘ	මුවා විසින්	මුවකු විසින්	මුවන් විසින්
කරණ	මුවාගෙන්	මුවකු ගෙන්	ඔුවන්ගෙන්
සම්ප්‍රධාන	මුවාට	මුවකුට	මුවන්ට
ඵවධි	මුවා කෙරෙන්	මුවකුගෙන්	ඔුවන් කෙරෙන්
සම්බන්ධ	මුවාගේ	මුවකුගේ	මුවන්ගේ
ආධාර	මුවා කෙරෙහි	මුවකු කෙරෙහි	මුවන් කෙරෙහි
ආලපන	මුවා		මුවනි

Table 1: Conjugation table for the Sinhala word 'මුවා'

To implement the each "Gana" (Conjugation tables) in Sinhala, we used the palindrome table for the each noun group. The table 2 shows the palindrome table for the 'මුව ගණය. This table gives the way to generate each noun forms through the add and/or remove letters form it based form.

Word form	add	remove	example
Base form	මුව		
Direct Singular	ා		මුවා
Indirect subject	වෙක්	ව	මුවෙක්
Indirect object	කු		මුවකු
Plural subject	වෝ	ව	ඔුවෝ
Plural subject	න්		මුවන්

Table 2: Palindrome table for the Sinhala 'මුව ගණය'

No.	Case විභක්තිය	ීමසෙං
1	Nominative (ප්‍රථමා)	-
2	Accusative (කර්ම)	-
3	Instrumental (කතෘ)	විසින්
4	Auxiliary (කරණ)	ගෙන්
5	Dative (සම්ප්‍රධාන)	ට/ හට
6	Ablative (අවධි)	ගෙන්
7	Genitive (සම්බන්ධ)	ගේ
8	Locative (ආධාර)	කෙරෙහි
9	Vocative (ආලපන)	-

Table 3: Sinhala case suffixes

In addition to these rules, the case rules are used to create case effect of a noun. The table 3 shows some suffixes for the case rules.

Sinhala Verb Morphology

Sinhala verbs are divided into two general classes, namely, transitive verb (sakarmaka) and interansitive (Akarmaka). Further, these two verb categories are inflected for voice (karaka) , mood (vidi) , tense (kala), number (wachana) and person (purushaya). Voice can be either active or passive. There are four types of moods, namely, indicative, optative, imperative and conditional[22]. Sinhala language has only three tenses. They are Past tense, Present tense and future tense. Main verb (Akkyathaya) praticipate three types of inflections namely person, number and sex.

Furthermore, structures of the Sinhala verbs are different from English. In comparison with the English language, the Sinhala language has only three tenses such as present (Varthmana), past (Athitha) and future (Anagatha) and the English shows 20 tenses for active and passive. Note that, More than 18 inflection forms are available in a Sinhala base verb including inflection of the tense, number and the person. Also there are four moods such as Indicative mood, Optative mood, Imperative mood and Conditional Mood and two participles Present participle (Misrakriya) and Past participle (Purvakriya). For example යමින්, ගොස් is the inflection form of the above two participles. The table 4 shows some inflection forms of the Sinhala verbs.

Person	**Number**	**Present**	**Past**	**Future**
First	Singular	කමි	කන්නෙමි	කෑවෙමි
First	Plural	කමු	කන්නෙමු	කෑවෙමු
Second	Singular	කහි	කන්නෙහි	කෑවෙහි
Second	Plural	කහු	කන්නෙහු	කෑවෙහු
Third	Singular	කයි	කන්නේය	කෑවේය
Third	Plural	කති	කන්නෝය	කෑවෝය

Table 4: inflection form of the Sinhala verb

From the morphological point of view, a verb contains two parts, namely, Base verb and a suffix. Base verb is a prakurthi, and it is named as kriya prakurthi. Diffrent verb forms are generated by adding diffrent suffixes for the kriya prakurthi.

In addition to the above, other parts of speech namely Nipatha and Upasarga do not participate any inflections.

III OVERVIEW OF THE ENGLISH TO SINHALA MACHINE TRANSLATION SYSTEM

The English to Sinhala Machine translation system (BEES) has been develop as Multi Agent System (MAS)[8]. The Multi agent system [5][24] contains nine agents, namely, English Morphological Agent, English Parser Agent, English to Sinhala base word translator agent, Transliteration Agent, Sinhala morphological generator agent, Sinhala Parser Agent, Intermediate editor agent, Message space agent and the Request agent. Also system uses four lexical dictionaries namely, English dictionary, Sinhala dictionary, English to Sinhala bilingual dictionary and concept dictionary as a common ontology of the Multi agent system. Figure 1 shows design of the BEES.

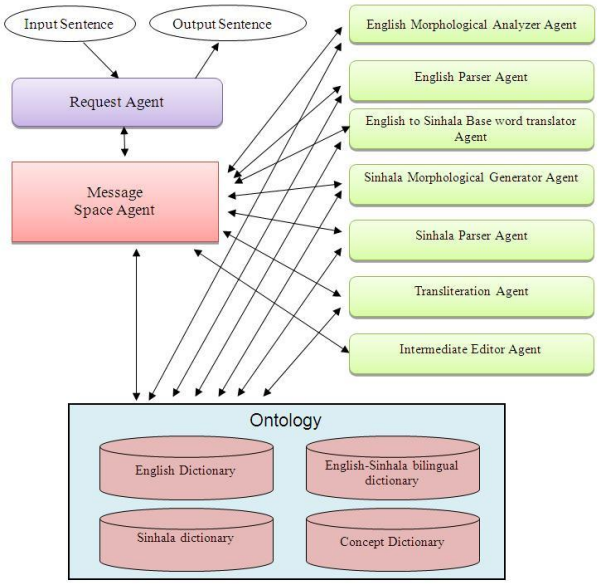

Figure 1: Design of the BEES

A. *English Morphological Agent*

English Morphological agent access English morphological analyzer and reads the result of the morphological analysis. These agents are automatically created for each word in the input sentence. Using English morphological analyzer English Morphological agent analyze an English word. After analyzing the English word agents are self-destroyed from the system. English Morphological agents are implemented by using Java.

B. *English Parser Agent*

After English Morphological agents analyzing the all words in the input sentence English parser agent analyzes the input sentence by using English parser. The English parser agent reads source English sentence and results from English Morphological agents. Then parser agent accesses the English parser and reads the result of the Syntax analysis. The English parser works as a syntax analyzer [9]. The parser has also been implemented using SWI-PROLOG. The English parser agent has been developed using JAVA.

C. *English to Sinhala base word Translator Agent*

The English to Sinhala based word translator agent uses base-word translator to translate English base-word into Sinhala base-word. Bilingual translator agents are automatically created for a given English word and identify the most suitable Sinhala based-word. The base-word translator translates English base word[25] into Sinhala base word by using concept dictionary[7] and the English to Sinhala bilingual dictionary. The Bilingual translator agent read this translation result from word level translator and negotiated from each other agents. Then Bilingual translator agent identifies the most suitable Sinhala based-word for a given English word. The Bilingual translator agent has also implemented using JAVA.

D. *Sinhala Morphological Generator Agent*

The Sinhala Morphological Agent access Sinhala Morphological analyzer [10] (Sinhala word generator) and reads the output result of the Sinhala Morphological generation. Sinhala morphological agents are created for each Sinhala base-word which is provided by the English to Sinhala based word translator agent. Figure 2 shows the design of the Sinhala morphological agent with Sinhala Morphological generator and the Sinhala dictionary

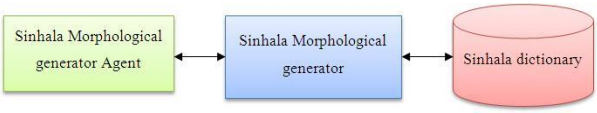

Figure 2: Design architecture of the Sinhala Morphological generator agent

The Sinhala morphological generator is the key module of the English to Sinhala machine translation system. This generator fundamentally works through the concepts of Varanegeema in Sinhala language. The Concepts of Varanegeema is the theoretical basic of the English to Sinhala machine translation. The Sinhala morphological generator accesses the Sinhala dictionary and generates appropriate Sinhala word forms. This morphological generator has been designed through the Sinhala grammar rules to generate appropriate Sinhala words. To generate a Sinhala noun, it requires base-word, number, case and the form of the noun (direct or indirect). The Sinhala verb requires base-word tense, person and number. The Sinhala adjective requires type and the base-word. The adjective type may be positive, comparative or superlative. The other type of words do

not participate the word conjugation. Therefore these words are stores in the Sinhala dictionary and the generator reads these words directly.
After generating the suitable Sinhala word each agents are automatically destroyed from the system. The Morphological agent has also implemented using JAVA and Sinhala Morphological analyzer implemented using SWI-PROLOG.

A. Sinhala Parser Agent

The Sinhala parser agent (SPA) accesses the Sinhala parser [12] and generates the corresponding Sinhala sentence by using Sinhala parser. The Sinhala parser works as a Sinhala sentence composer. It receives Sinhala words from the Sinhala morphological agent and composes grammatically correct Sinhala sentence. The Sinhala parser has also implemented using SWI-Prolog. The Sinhala parser agent has been developed using JAVA.

B. Transliteration Agent

Translation system needs to solve Out-of-vocabulary problems and handle technical terms. Machine transliteration can be used as a resalable solution for that. At present we have developed two types of transliteration models [12]. One of these models transliterates Original English text into Sinhala Transliteration and the other transliterate Sinhala words that are written in English which transliterate into Sinhala. Finite State transducers are used to develop these two modules. By using this prolog based transliteration modules we have been implemented a transliteration agent. This agent can transliterate English text into Sinhala. If English Morphological agent cannot be analyzed a given word then transliteration agent transliterates it. Transliteration agent is developed by using JAVA.

C. Message space Agent

The Manager agent is the controlling agent of the system. This agent handles all the process of the other agents. Major task of the Messaging is provides messages for each agent and monitor the translation process. The Message agent has been developed using java.

D. Request Agent

The Request agent is create a bob for translation and get the output result from the system The Message agent has been developed using java.

E. Intermediate editor agent

The Intermediate editor agent is used to facilities to edit the text by using human support through the intermediate editor[7]. The intermediate editor provides facilities such as showing synonyms, anti-synonyms, related words, etc. The intermediate-editor is linked up both English and Sinhala dictionaries in the MT system.

F. Lexical dictionaries (Ontology)

The Translation system uses four dictionaries such as English dictionary, English-Sinhala bilingual dictionary, Sinhala dictionary and Sinhala concept dictionary [9][11]. English dictionary contains English words and the lexical information. English to Sinhala bilingual dictionary is used to identify appropriate Sinhala base word for a given English word and it contains relation between English and Sinhala words. Sinhala dictionary stores Sinhala regular base words, irregular words and lexical information with set of rules for words conjugation. The concept dictionary is used to store concepts information to handle semantics.

IV SINHALA WORD GENERATION

This section describes how Sinhala morphological generator generates Sinhala words such as Nouns, verbs, adjectives, adverbs and prepositions through the concept of Varanegeema. This Sinhala morphological generator has been implemented through the SWI-Prolog. To generate appropriate Sinhala Nouns and Verbs the Sinhala morphological generator uses the Sinhala dictionary with set of rules. These rules are implemented based on the Sinhala grammar. The following two sub sections reports how Sinhala morphological generator generates the Sinhala noun and verbs.

A. Sinhala Noun generation

To generate a Sinhala noun the generator uses a base word id of the particular Sinhala noun, case and direct or indirect form. The system uses the Prolog predicate named get_sin_noun/8 to generate Sinhala noun. For example the following predicate generate the Sinhala noun with the based id ‘s0001’ and Singular, direct subject form of the Sinhala noun the parameter sg is refers to Singular, dr is refers to direct and the v1 is refers to the subjective case of the Sinhala Noun.

```
get_sin_noun(s0001,P,sg,S,L,dr,v1,OUT).
```

After generating the Sinhala noun system returns the output word through the parameter named OUT. Also it gives relevant grammar information such as Person, Sex, Live for the given noun. Note that, the noun generation process is done through the two stages; As the first step the system check whether the noun is in the irregular form. If it is available on the English dictionary then system directly uses it. These irregular nouns are stored in the prolog database on the Sinhala dictionary by using sn/8 prolog predicate. The

following rule shows how generator gets the irregular nouns form Sinhala dictionary.

```
get_sin_noun(WID,P, N,S, L, DI, VB, NW) :-
 sn([WID], P, N, S, L, DI, VB, NW).
```

If there is no irregular form available, then system searches the regular nouns. To generate suitable noun form Sinhala morphological generator uses Sinhala regular noun information which is available in the sn/9 predicate format on the Sinhala dictionary. This sn/9 predicate gives Person (PE), Sex (S), Live(L), noun conjugation id(NID) and case rules for three noun forms Singular direct, Singular indirect and plural.
`sn([WID],PE,S,L,NID,C1,C2,C3,WD).`
These rules (noun forms) are based on the palindrome tables for the Sinhala noun. The following code shows how prolog represent the Palindrome for Sinhala noun (;dr .Kh) The noun_posfix/4 predicate gives noun conjugation id, live, form add part and the remove part of the word.

noun_posfix(s933001, li, bas, '', 'ව').
noun_posfix(s933001, li, sds, 'ව', '').
noun_posfix(s933001, li, sdo, 'ව', '').
noun_posfix(s933001, li, sis, 'වෙක්', '').
noun_posfix(s933001, li, sio, 'වෙකු', '').
noun_posfix(s933001, li, pds, 'වෝ', '').
noun_posfix(s933001, li, pdo, 'වන්', '').

After that by using noun conjugation id the system get the noun generation methods. Using this information system generates the suitable Sinhala noun. This process uses three steps namely base word generation, word form generation through the base form and finally gets the correct Sinhala word through the adding the case suffix. The following sample code shows some case suffix which are used the word generation.

noun_vib_postfix(s910000, v1, '', '').
noun_vib_postfix(s910000, v2, '', '').
noun_vib_postfix(s910000, v3, ' විසින්', '').
noun_vib_postfix(s910000, v4, 'ගෙන්', '').
noun_vib_postfix(s910000, v5, 'ට', '').
noun_vib_postfix(s910000, v6, 'ගෙන්', '').
noun_vib_postfix(s910000, v7, 'ගේ', '').
noun_vib_postfix(s910000, v8, ' කෙරෙහි', '').
noun_vib_postfix(s910000, v9, ',').

A. Sinhala Verb Generation

Sinhala verb generation process is same as the Sinhala word generation. Sinhala verb generation is done through the two steps. As a first step the Sinhala morphological generator search required verb is available as the irregular form. If it is available on the Sinhala dictionary then system directly uses it. These irregular verbs are stored in the prolog database on the Sinhala dictionary by using sfv/6 prolog predicate. The following rule shows how generator gets the irregular nouns form Sinhala dictionary.

```
get_sin_final_verb(S, T, P, N,Te, SW):-
 sfv([S], P, N, T, Te, SW).
```

The following code shows how system stores the Sinhala irregular verbs in the Sinhala dictionary.

sfv([8000002], fs, sg, at, pr, 'සිටිමි').
sfv([8000002], fs, pr, at, pr, 'සිටිමු').

If there is no irregular form available, then system searches the regular verbs. To generate suitable verb form Sinhala morphological generator uses Sinhala regular verb information which is available in the sfv/9 predicate format on the Sinhala dictionary. This sfv/9 predicate gives all the rule ids to generate verb forms.
`sfv([WID],R1,R2,R3,C1,C2,C3,O1,WD).`
These rules (verb forms) are based on the palindrome tables for the Sinhala verb. The following code shows how prolog represents the Palindrome for Sinhala verb

verb_posfix(s910001, at, fs, sg, pr, 'මි', 'නවා').
verb_posfix(s910001, at, fs, pr, pr, 'මු', 'නවා').
verb_posfix(s910001, at, sc, sg, pr, 'හි', 'නවා').
verb_posfix(s910001, at, sc, pr, pr, 'හු', 'නවා').
verb_posfix(s910001, at, td, sg, pr, 'යි', 'නවා').
verb_posfix(s910001, at, td, pr, pr, 'ති', 'නවා').

The verb_posfix/4 predicate gives verb conjugation id, voice, person, number, tense and add part and the remove part of the word. After that by using verb conjugation id the system generates the corresponding verb for the given grammar information.
Note that, these Sinhala word generation procedures are based on the Sinhala grammar. In Addition to the noun and verb generation all other words are available in the irregular forms.

V. EVALUATION

The English to Sinhala Machine Translation system has been evaluated through the following three stages;

1. Conducted a white box testing approach and tested each module in the Machine Translation system through the developed testing tools (module testing)
2. Evaluated the system performance and calculated the error rate through the evaluation test bed (performance testing)
3. Intelligibility and accuracy test was conducted through the human support. (Accuracy testing)

The Sinhala morphological generator has been tested through the white box testing approach. To test the

Sinhala word generation we have created a testing tool. This tool gives all the word form of the given Sinhala based word. For example this tool provides 27 word form for a Sinhala noun and the 38 forms for a base verb. To test each rules we have used randomly selected 300 Sinhala words and tested each word through the testing tool. The experimental result shows 92 % accuracy for the Sinhala word generation. Very little number of word generations are shows incorrect generation. This is because there forms have some irregular patterns. After updating the irregular patterns it shows 100 % accuracy for the given 300 words. The figure 3 shows how testing tools give Sinhala word forms in the prolog shell.

In addition to above, word error rate (WER) has been calculated through the selected 200 sentences. The word error rate is calculated through the following formula.

$$Word\ Error\ Rate = \frac{\sum \text{Number of incorrect word in the sentence}}{\sum \text{Number of words in the sentence}}$$

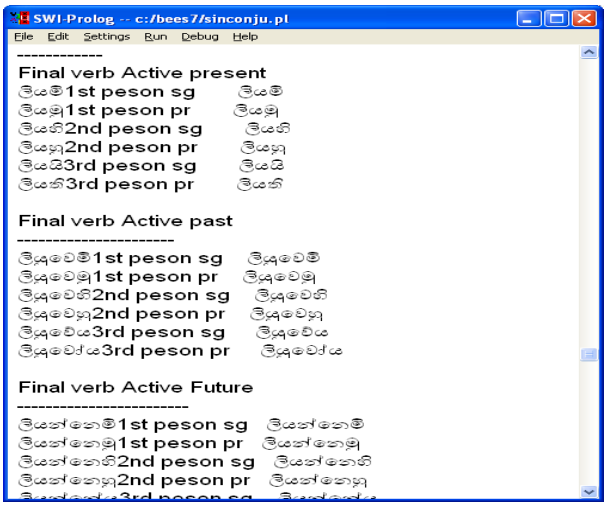

Figure 3: how testing tool works in the prolog shell

The following example sentence shows how English to Sinhala machine translation system gives the Sinhala translation for the given English sentences.

The good boy writes a new book
හොඳ පිරිමි ළමයා අලුත් පොතක් ලියයි

I gave a book to the good boy
මම පොතක් හොඳ පිරිමි ළමයාට දුන්නෙමි

She gave a new book to my good friend
ඇය අලුත් පොතක් මගේ හොඳ මිතුරාට දුන්නේය

A good boy is writing a new book
හොඳ පිරිමි ළමයෙක් අලුත් පොතක් ලියමින් සිටියි

I will be writing a new book
මම අලුත් පොතක් ලියමින් සිටින්නෙමි

The above example sentence shows the English terms 'the boy', 'A boy', and 'to the good boy' are generate as a 'පිරිමි ළමයා' 'පිරිමි ළමයෙක්', and 'හොඳ පිරිමි ළමයාට'. These word form generation is done through the Sinhala morphological generator.

The table 4 shows final evaluation results of the English to Sinhala machine translation.

Evaluation Results	Percentage
Word error rate (WER)	7.2 %
Sentence error rate (SER)	5.4 %
Intelligibility and accuracy	89.1 %

Table 4: Final evaluation results

The figure 4 shows user interface of the English to Sinhala machine translation system with some translation results.

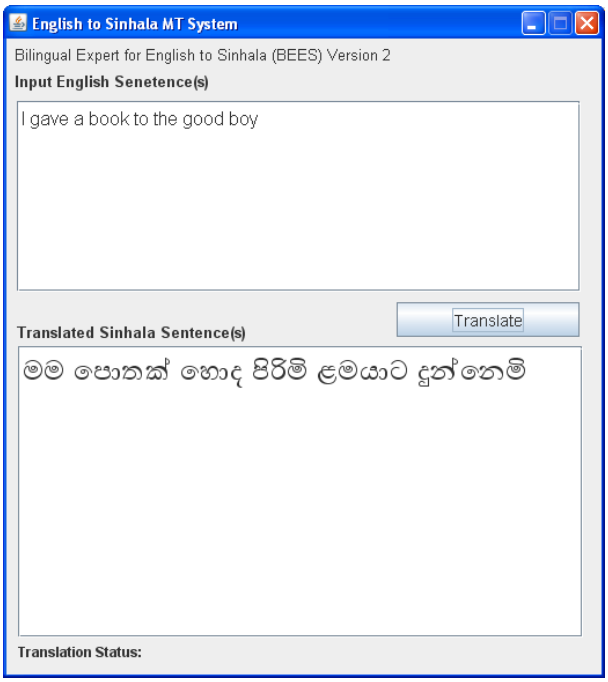

Figure 4: User interface of the Enflish to Sinhala machine translation system

VI. Conclusion and Further Works

This paper has reported detail of the concept of Varanegeema. The concept of Varanegeema reduces 93% of the Sinhala noun availability on the Sinhala dictionary as well as 97% of the Sinhala verb availability. This is because, the concept of Varanegeema helps to generate more than 27 nouns from for a base noun and more than 36 verbs forms for

a base verb. In addition to the above Sinhala morphological generator can directly use for the several Sinhala language applications such as Sinhala spell checker, Sinhala grammar checker etc.

Design, implementation and evaluation methodology of the English to Sinhala machine translation system also discussed in the paper. The experimental result shows 89 % accuracy of the system and 7.2 % word error rate and the 5.4 % sentence error rate. At present system uses limited number of lexical resources. Updating lexical resources are the further work this project.

REFERENCES

[1] Akshar B., Chaitanya V., Sangal R., Natural Language Processing: A Paninian Perspective, Prentice Hall of India, New Delhi, India, 1995.

[2] Akshar Bharati, Vineet Chaitanya, Amba P. Kulkami, & Rajeev Sangal: Anusaaraka: machine translation in stages. Vivek: a Quarterly in Artificial Intelligence, vol.10, no.3, July 1997, pp.22-25.

[3] Akshar B., Sangal R., Sharma D. M., Mamidi R., Generic Morphological Analysis Shell, In Proceedings of LREC 2004 - SALTMIL Workshop: First Steps in Language Documentation for Minority Languages. Lisbon, Portugal, 2004.

[4] Mahesh R, Sinha K., Integrating CAT and MT in AnglaBhart-II architecture, 10th EAMT conference "Practical applications of machine translation", May 2005, pp235-244.

[5] Hettige B. Karunananda A.S., Multi-Agent architecture for English to Sinhala Machine Translation, Proceedings of 28th National Information Techonology conference (NITC10), Sri Lanka, September 2010

[6] B. Hettige, A.S. Karunananda, "Theoretical Based Approach to English to Sinhala Machine Translation ", Proceedings of the 4th International Conference on Industrial and Information Systems (ICIIS2009), IEEE, Sri Lanka, 2009.

[7] Hettige B. Karunananda A.S., Context-Based approach to semantic handling in English to Sinhala Machine Translation, Poster presentation on 27th National Information Techonology conference (NITC09), Sri Lanka, September 2009

[8] B. Hettige, A.S. Karunananda, "Swarm Intelligence of BEES for Machine Translation", Proceedings of ITRU Reserch symposium, University of Maratuwa, Sri Lanka 2009

[9] Hettige B., Karunananda A. S., A Parser for Sinhala Language – First Step Towards English to Sihala Machine Translation, proceedings of International Conference on Industrial and Information Systems (ICIIS2006), IEEE, Sri Lanka, 2006, pp583-587.

[10] Hettige B., Karunananda A. S., A Morphological analyzer to enable English to Sinhala Machine Translation, Proceedings of the 2nd International Conference on Information and Automation (ICIA2006), Colombo, Sri Lanka, 2006, pp. 21-26.

[11] Hettige B., Karunananda A. S., Developing Lexicon Databases for English to Sinhala Machine Translation, proceedings of second International Conference on Industrial and Information Systems (ICIIS2007), IEEE, Sri Lanka, 2007, pp. 215-220.

[12] Hettige B., Karunananda A. S., Transliteration System for English to Sinhala Machine Translation, proceedings of second International Conference on Industrial and Information Systems (ICIIS2007), IEEE, Sri Lanka, 2007, pp.209-214.

[13] Hemant D., Computer-assisted translation system – an Indian perspective, Machine Translation Summit VII, 13th-17th September 1999, Kent Ridge Digital Labs, Singapore, Proceedings of MT Summit VII "MT in the Great Translation Era"; pp.80-85.

[14] Wikipedia, http://www.wikipedia.org

[15] SWI-PROLOG home page; http://www.swi-prolog.org

[16] Weerasinghe R., "A Statistical Machine Translation approach to Sinhala-Tamil language translation, In SCALLA 2004.

[17] Hutchins, W. J., Machine Translation: A Brief History, Oxford Program Press, pp-431-445, 1995.

[18] Karurarathna, S., Sinhala Viharanaya, Wasana Prkashkaya, Dankotuwa, Sri Lanka. 2005.

[19] Karunathilaka, W. S., Sinhala Basha Viharanaya. Colombo 11, Sri Lanka: M. D. Gunasena nad Ltd., 2004.

[20] Disanayake, J. B., Basaka Mahima 6: Prakurthi. Colombo 10, Sri Lanka: S. Godage and Brothers., 2000.

[21] Disanayake, J. B., Basaka Mahima 5: Akuru ha Pili Colombo 10, Sri Lanka: S. Godage and Brothers., 2000.

[22] Gunasekera, A. M., A Comprehensive Grammar of the Sinhalese Language, New Delhi, India: AES Reprint. 1986.

[23] Girish N. J., The System of Panini, Language in india, Vol 4 Sep, 2004.

[24] George R., "A new direction of research into Artificial Intelligence", Sri Lanka Association for Artificial Intelligence 5th Annual Sessions, 2008.

[25] Wren P.C., Martin H., High School English grammar and Composition, S. Chand and Company Ltd, Ram Nagar, New Delhi, India, 2005.