A Translator from Sinhala to English and English to Sinhala (SEES)

L.Wijerathna#, W.L.S.L.Somaweera, S.L.Kaduruwana, Y.V.Wijesinghe, D.I.De Silva, K.Pulasinghe, S.Thellijjagoda

Sri Lanka Institute of Information Technology,
New Kandy Road, Malbe, Sri Lanka
#Email: *laksri.w@sliit.lk*

***Abstract*— This paper presents a rule based machine translation system which is capable of translating sentences from Sinhala to English and vice versa. This is the first Sinhala to English and English to Sinhala machine translation system which comes with features such as a Sinhalese font translator, which is capable of interpreting Sinhalese words written in English characters (Singlish) to Sinhala characters, and an English grammar and spell checker. An entered sentence to the system will be tokenized and translated according to a rule. When translating Sinhala sentences to English the user input should be in Singlish and when translating English sentences to Sinhala input should be in English. The main objective of this translator is to enable a smooth flow translation of words, sentences and paragraphs to locals as well as foreigners and thereby eliminate the language barrier. A considerable amount of rules, patterns and words of both languages were used to develop this system. With 87% accuracy this pilot machine translation system translated 500 grammatically well-structured Sinhala sentences to English and 150 grammatically well-structured English sentences to Sinhala. The system is capable of translating approximately 70 sentences in one minute.**

***Keywords*— Sinhala to English Translator, English to Sinhala Translator, English, Singlish, Sinhala, Rule base machine translation**

I. Introduction

Sinhala is a language that has originated from one of the branches of the Indo-Aryan group of languages. However, due to its exposure to other language families of the region it has some unique features that are not found in other Indo-Aryan languages [2]. The sound system of Sinhala consists of 14 vowels and 26 consonants [2]. In literary Sinhala which is written from left to right, the subject has to agree with the tense, gender and form (singular and plural) of the verb. Unlike English any written Sinhalese word is pronounced the same way it is written.

English is a West German language originating from England. It is the dominant language in the United Kingdom, United States, many Commonwealth nations including Australia, Canada, New Zealand and other former British colonies [6]. Due to its extensive use as an official language, English has now become the most important language around the world. Many people from non-English speaking countries tries their very best to learn this language as the written and spoken knowledge of English is considered as an additional advantage for them to improve in their professional and academic lives.

In machine language translations, translation of non-similar languages is much more difficult than the translation of similar languages. As an example, translating from English to French or German is easier because English, French and German have close relationship and similarities to each other since they are based on a similar alphabet [3].

Although many Sri Lankans are knowledgeable enough to work in the Sinhala language most of them do not have sufficient knowledge to work in English. On the other hand, most of the foreigners who visit Sri Lanka do not have any understanding about the Sinhala Language. Thus, this language translator was build with the intension of overcoming this barrier up to some extent.

In this paper we present a, Sinhala to English and English to Sinhala language translator for both active and passive voices, eliminating the limitations that exist in the currently available translators. In addition to the translator SEES consists of Sinhalese font translator, which is capable of interpreting Singlish words in Sinhala characters, and an English grammar and spell checker. Rather than doing a word by word translation the system does a sentence vice translation in order to give a more meaningful translation.

The system SEES use tokenization in advance to increase the performance and accuracy of the translation. Tokenization is concept of identifying the constructive segments of the sentence and categorizing them accordingly [4]. Constructive segments are translated and rearranged according to the rules of the translated language with the use of a rule base engine.

The rule base engine is responsible of generating the English translation according to the input SinGlish translation. In order to do this rule base engine is associated with a Knowledge base. Previously identified grammar rules are presented in this knowledge base. When the tokens are identified it will check with the available rules and then will be translated according to the best fit of the rule which is given in the knowledge base.

To increase the user friendliness the input sentences which the user enters in Singlish is translated into its corresponding Sinhala font.

The pilot system is developed to achieve high performance with the use of parallel processing and to increase the accuracy of the translation. Parallel processing is achieved with the use of threading. So the pilot system uses more than one thread of execution to achieve the parallel processing. It is capable of handling large sentences with adjectives, adverbs, subjects, objects, mix verbs, pre verbs, verbs, conjunctions, articles, prepositions etc. [5].

II. Literature Review

In the modern era machine translation has received much attention due the decrement in time and human effort required for converting and translating words, sentences and paragraphs. Thus, many translators have been proposed for different languages. Most of them have been focused on Indo-European, Indo-Aryan or Sino-Tibetan families [1].

As the language structures of Sinhala and English languages are completely different from one another they are categorized as non-similar languages. Due to this only a few translator systems have been proposed for Sinhala to English or vice versa.

978-1-4673-5530-8/12/$31.00 ©2012 IEEE

Source [12] cites an example based machine translation system for English to Sinhala translations. It was developed mainly to be used in the government domain. It achieved scores of 0.17 - 0.26 for Bleu translation metric [7] for 3-gram analysis using one reference translation. The English to Sinhala translator proposed in [13] achieved 89% accuracy with a word error rate of 7.2% and a sentence rate error of 5.4% for two hundred sample sentences. A translator from Sinhala to English is proposed in [3]. For 150 grammatically well-balanced active voice Sinhala sentences it achieved 75% accuracy.

Convertors from Singlish to Sinhala have also been proposed. Some of the available convertors are the Unicode Converter [8], the Google Transliteration IME [9] and the convertor found in [10]. All three of them allow users to type a word online according to the way that it sounds in English (not to its meaning) and get the corresponding Sinhala word with correct Sinhala fonts [8].

Dictionaries which are capable of giving the English meaning of any word typed in Singlish and the Sinhala meaning of any word typed in English have also been proposed. Madura dictionary [11] is one such dictionary storing over 230,000 English and Sinhala definitions.

III. Methodology

Object oriented methodology was used to design the system.

The system uses encapsulation to hide the complexity of the system from the users. The entire tasks are done behalf of the user by single button click. What the user wants to know is what is the input format that user should enter and which button should be clicked to get the output. Separate classes are created for each entity and the methods are created to perform the particular actions. So the structure of the system is well organized where user actions can be handled by the object creation and the method invocation.

The work was carried out using the Visual Studio .NET environment with C#.Net as the programming language. The knowledge bases were developed in SQL 2005 and Linq was used for the connectivity of the database. SEES consists of two main components.

1. Sinhala to English Translator
2. English to Sinhala translator

A. Sinhala to English Translation Component

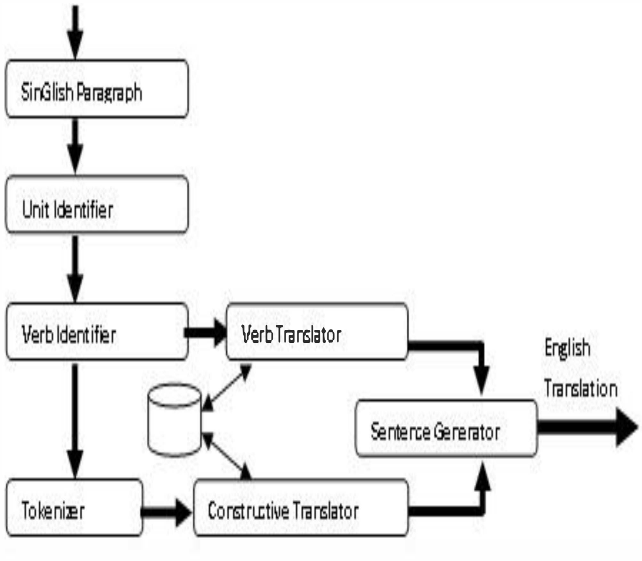


Fig. 1 Behavior of Sinhala to English Translator

Fig. 1 shows the system diagram of the Sinhala to English translator. The task of this component is to translate the Singlish input into its corresponding English translation. First from the user input sentence/unit is identified. The unit is identified using the full stop as a delimiter. So the pilot system only accepted the sentences that are ended with the full stop. System SEES processes one unit (sentence) at a time.

The pilot system is tokenizing the identified unit as the next step of the process. In the tokenizing process first the constructives are identified. Each constructive is identify with the use of the space as a delimiter. In the tokenization constructive is categorized as a verb, subject, noun, adjective, adverb or article. This is done by parsing the understood constructive to the knowledge base. When the constructive is parsed, by looking at the available knowledge in the knowledge base system tokenizes them accordingly.

Following example shows how the tokenizing process is done for a simple sentence to identify its constructive segments. In this example the user input is “sudhu lamaya: we:gayen: gedara dhiw:we:ya.” User has to enter the text according to the system defined standards. As an example the word “සුදු” should be typed as “sudhu”. System expects a space between each word.

sudhu	lamaya:	we:gayen:	gedara	dhiw:we:ya.
සුදු	ළමයා	වේගයෙන්	ගෙදර	දිව්වේය.
<adj>	<sub>	<adv>	<noun>	<vrb>

From a unit, verb is identified first. The identified verb is then stemmed (split in to base and postfix). Fig. 2 shows how the stemming is applied for a verb to identify the base and the postfix. To increase the efficiency and performance of the translation only the base part of the verb was stored in the data base.

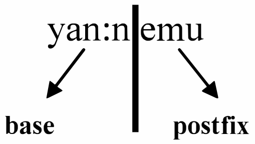


Fig. 2 Verb Stemming

A verb has different variations according to the tense. As an example the verb ‘යන්නෙමි’ (yan:nemi) have the following variations:

- යන්නෙමු (yan:nemu)
- යන්නේය (yan:ne:ya)
- යන්නීය (yan:ni:ya)
- යන්නෝය (yan:no:ya)
- යන්නෙහි (yan:nehi)
- යන්නෙහු (yan:nehu)

Simultaneously the remaining constructive segments are tokenized accordingly with the use of a knowledge base. Next, the identified tokens are translated with the use of a knowledge base and generate the final grammatically correct English output sentence with the use of a rule base engine. When generating the final translation, with the use of the sequence identifying algorithm the appropriate rule is identified and according to the identified rule the translated sentence is rebuild to make it meaningful.

The translations of the identified tokens are done with the use of a threaded allocation to achieve the parallel processing which help to increase the performance of the system.

In the Sinhala language, passive voice sentences belong to different forms and parts of the speech. In addition to translating the active voice forms, SEES is able to handle passive voice translations with the use of tokenization up to some extent. Passive voice acts as reverse of active voice. In the translation process, object becomes the subject. There the verb is always kept according to the object which is known as "Karma Karaka" in Sinhala language. Following is an example of the tokenization process for a Sinhala passive voice sentence.

ඔවුන් විසින් මා දකිනු ලබමි.

<Sub> <Prep> <Obj> <Verb>

Specific rules are applied for the translation process and then appropriate English words are retrieved using the relevant table in the database. Then the relevant pattern of the sentence is identified. Next the extra words are added in order to make the sentence meaningful and finally the sentence generator generates the final output as shown below.

I am seen by them.

<Obj> <Verb> <Prep> <Sub>

When finalizing the words the system checks for the proper usage of articles and prepositions. For that it identifies the postfix of the noun as shown in the Table I and according to the postfix it adds the required article or the preposition.

TABLE I
ARTICLES AND PREPOSITION TRANSLATION

Postfix	Example	Article	English Translation
k: (ක්)	pothak: (පොතක්)	a	a book
ta (ට)	gedharata (ගෙදරට)	to	to home
dhi: (දී)	ka:maraye:dhi: (කාමරයේදී)	in	in room
yen: (යෙන්)	kadayen: (කඩයෙන්)	from	from shop
sin: (සින්)	gasin: (ගසින්)	from	from tree

The system is capable of translating the sentences with mix verbs, pre verbs and dummy verbs. First with the help of tokenization system identifies the verb type and then does the translation accordingly. Next it adds the relevant words to the translated sentence in order to make it more meaningful. Table II shows the different type of verbs and how they are translated by the system.

TABLE II
VERB TYPES and TRANSLATIONS

Example	Verb Type	Translation	Additional Words
natamin: (නටමින්)	mix verb	dancing	while
kiyawa: (කියවා)	pre verb	read	-
pa:dam: karan:ne:ya (පාඩම් කරන්නේය)	dummy verb	studies	
sel:lam: karamin: (සෙල්ලම් කරමින්)	dummy verb + mix verb	playing	while

B. *English to Sinhala Translation Component*

This component translates Simple Present Tense, Simple Past Tense and Simple Future Tense English sentences into Sinhala. The system initially tokenizes the entered English sentences as follows:

I went home.
<Sub> <Verb> <Noun>

Then the verb of the sentence is indentified. Sinhala translation of a verb has two parts namely prefix and postfix as shown in fig. 3. Postfix of a verb may differ according the person type, gender and singular and plural forms. As an example the verb '**went**' can be translated in Sinhala as:

- giyemi (ගියෙමි) - First person , Singular
- giyemu (ගියෙමු) - First person, Plural
- giya:ya (ගියාය) - Third person, Female, Singular
- giyo:ya (ගියෝය) - Third person, Male, Plural etc.

The prefix of the verb translation is retrieved from the database and postfix is added from the system. The same procedure is used when storing and translating the nouns as well. That is because, when there are prepositions in front of a noun, the translation of a noun may also vary. The translation of the noun '**tree**' with and without prepositions is as follows:

- tree - gasa (ගස)
- a tree - gasak (ගසක්)
- from a tree - gasen (ගසෙන්) etc.

When translating an English sentence into Sinhala the verb is placed as the last word of the sentence as the following example. Ex.: I **went** home. → මම ගෙදර **ගියෙමි.** ("mama gedara **giyemi**")

Adjectives and adverbs are placed before their relevant noun or verb as the following example. Ex.: **Small** child went home **quickly.** → පොඩි ළමයා ගෙදර වේගයෙන් ගියේය.

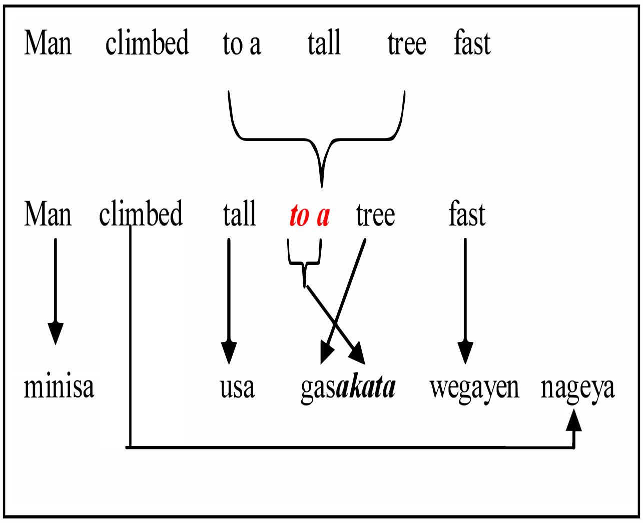


Fig. 3 Demonstration of how words are arranged in a Sinhala translation for a given English input.

IV. Research Findings and Evidence

A machine does not have the human dimension required for dealing accurately with the subtle nuances of a natural language. Thus, no machine translator has been capable of translating a sentence given in Sinhala to English or vice versa with 100% accuracy. However, with 87% accuracy this pilot machine translation system translated 500 grammatically well-balanced Sinhala sentences to English and 150 grammatically well-structured English sentences to Sinhala. It is able to translate sentences in any tense: present, past and future, for both singular and plural forms, all the three person types: first person, second person and third person and all the genders. The following sub sections describe some sentence patterns that were correctly translated by SEES.

A. Sinhala Words with Multiple English Meanings

In language translation some words of source language can provide multiple meanings for the targeted language. For example as illustrated in fig. 4 a sentence can contain a verb which provides multiple meanings when translated. The system is capable of identifying such occurrences. In such situations SEES generates all the possible outputs and gives the user the opportunity to select the most suitable translation.

Input	Lamaya: pa:ra **pAn:ne:ya** ළමයා පාර **පැන්නේය.**
Output 1	Child **crossed** the road
Output 2	Child **jumped** the road

Fig. 4 A Sinhala Word with Multiple English Meanings.

B. Sinhala Words with Multiple English Words

For some words in the Sinhala language there are many words in the English language. Fig. 8 demonstrates an example for such a sentence.

Input	**sudhu** Lamaya: pothak: kiyawan:ne:ya **සුදු** ළමයා පොතක් කියවන්නේය.
Output 1	**White** child reads a book.
Output 2	**Fair** child reads a book.

Fig. 5 A Sinhala Word with Multiple English Words.

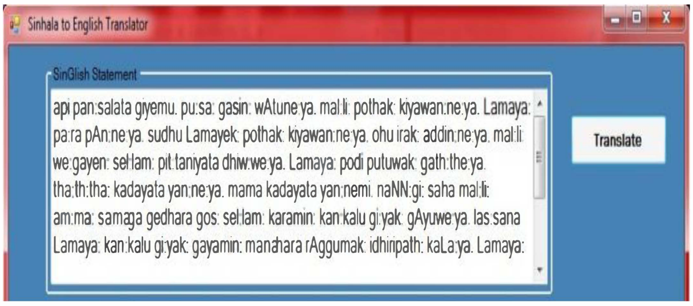
C. Ambiguous Words of Sinhala Language

One of the main barriers in language translation is ambiguity of natural languages. Some Sinhala words are ambiguous and can only be identified by the human brain. The meaning of the word depends on the place where the particular word is used. Fig. 6 demonstrates an example for a sentence with an ambiguous word. In such situations also the system generates all the possible outputs and gives the user the opportunity to select the most suitable translation

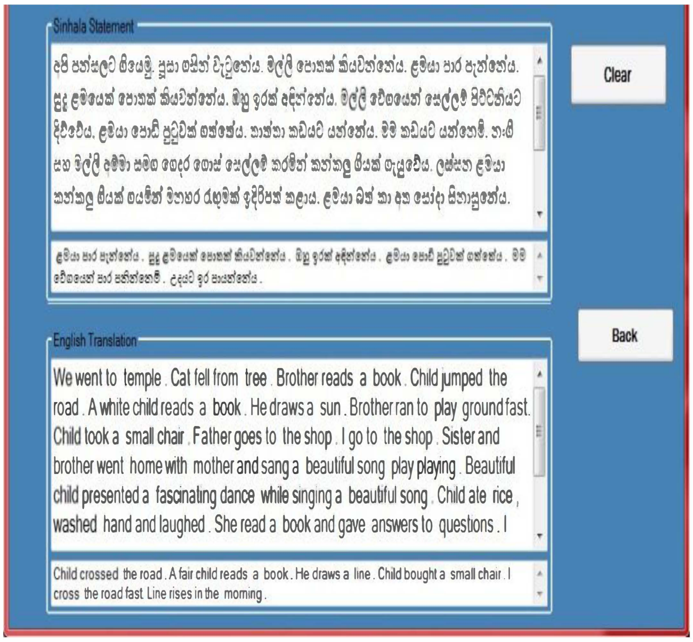
Input	ඔහු ඉරක් අඳින්නේය.
Output 1	He draws **a line.**
Output 2	He draws **a sun.**

Fig. 6 Ambiguous words of Sinhala language

Fig. 7 illustrates the interface of the Sinhala to English component of SEES. First text box is used to enter the Singlish words. According to the entered Singlish words Sinhala words will be generated in the second text box. Out of those words Sinhala words with multiple English meanings, Sinhala words with multiple English words and ambiguous words of Sinhala language would be separately identified and displayed in the third text box. Fourth text box displays the corresponding English translations for the entered Sinhala sentences. Fifth text box display the possible sentences for sentences in the third text box.



api pan:salata giyemu. pu:sa: gasin: wAtune:ya. mal:li: pothak: kiyawan:ne:ya. Lamaya: pa:ra pAn:ne:ya. sudhu Lamayek: pothak: kiyawan:ne:ya. ohu irak: addin:ne:ya. mal:li: we:gayen: sel:lam: pit:taniyata dhiw:we:ya. Lamaya: podi putuwak: gath:the:ya. tha:th:tha: kadayata yan:ne:ya. mama kadayata yan:nemi. naNN:gi: saha mal:li: am:ma: samaga gedhara gos: sel:lam: karamin: kan:kalu gi:yak: gAyuwe:ya. las:sana Lamaya: kan:kalu gi:yak: gayamin: manahara rAggumak: idhiripath: kaLa:ya. Lamaya:



අපි පන්සලට ගියෙමු. පූසා ගසින් වැටුනේය. මල්ලී පොතක් කියවන්නේය. ළමයා පාර පැන්නේය. සුදු ළමයෙක් පොතක් කියවන්නේය. ඔහු ඉරක් අඳින්නේය. මල්ලී වේගයෙන් සෙල්ලම් පිට්ටනියට දිව්වේය. ළමයා පොඩි පුටුවක් ගත්තේය. තාත්තා කඩයට යන්නේය. මම කඩයට යන්නෙමි. නංගී සහ මල්ලී අම්මා සමග ගෙදර ගොස් සෙල්ලම් කරමින් කන්කලු ගීයක් ගැයුවේය. ලස්සන ළමයා කන්කලු ගීයක් ගයමින් මනහර රඟුමක් ඉදිරිපත් කළාය. ළමයා බත් කා අත සෝදා සිනාසුනේය.

ළමයා පාර පැන්නේය . සුදු ළමයෙක් පොතක් කියවන්නේය . ඔහු ඉරක් අඳින්නේය . ළමයා පොඩි පුටුවක් ගත්තේය . මම වේගයෙන් පාර පනින්නෙමි . උදයට ඉර පායන්නේය .

We went to temple . Cat fell from tree . Brother reads a book . Child jumped the road . A white child reads a book . He draws a sun . Brother ran to play ground fast. Child took a small chair . Father goes to the shop . I go to the shop . Sister and brother went home with mother and sang a beautiful song play playing . Beautiful child presented a fascinating dance while singing a beautiful song . Child ate rice , washed hand and laughed . She read a book and gave answers to questions . I

Child crossed the road . A fair child reads a book . He draws a line . Child bought a small chair . I cross the road fast Line rises in the morning .

Fig. 7 System interface of Sinhala to English translator with examples

V. Conclusions and Future work

This paper presents a way of how machine translation can be used to help remove the language barrier between English and Sinhala languages.

The system consist of a Sinhala to English language translator which is capable of translating sentences of both active and passive voices, a Sinhalese font translator, which is capable of interpreting Singlish words in Sinhala characters, an English to Sinhala translator which is capable of translating English sentences in simple present tense, simple past tense and simple future tense, and an English grammar and spell checker. At present, the system has achieved a success rate of 87% with a corpus of 500 Sinhala sentences and 150 English sentences. SEES is capable of translating approximately 70 well-structured sentences in one minute. The system covers most of the complex areas in translations like sentences with mix verb, pre verb, and dummy verb and able to translate the sentences with proper usage of prepositions, articles, conjunctions, adjectives and adverbs. Furthermore it is capable of translating complex large sentences such as sentences with 12 words.

One of the main barriers in further enhancing the translator is the word ambiguity. Some words cannot be translated without proper human interaction. As a future work system can be developed using statistical base analysis with the involvement of a trained corpus.

The other remaining work with regards to the translator is to add more words to the knowledge base. Currently translated output is generated with the use of a rule base where only the introduced rules are being used. Our focus is to automate the currently used process with the use of artificial intelligence.

Acknowledgment

We would like to express our utmost gratitude to our parents who have contributed in innumerable ways towards the success of this project and have always been a source of inspiration.

References

[1] Axis Translation s Ltd. (2011) Indo-European languages. [Online]. Available: http://www.axistranslations.com/translation-article/indo-european-languages.html

[2] E. Perera (2000) Some unique features of Sinhala Language [Online]. Available: http://www.lankalibrary.com/

[3] D. De. Silva et al, "Sinhala to English Language Translator, " in *Proc. 4th International Conference on Information and Automation for Sustainability (ICIAFS)*, Dec. 2008, Sri Lanka, p 419 -424.

[4] R. Weerasinghe, "Bootstrapping the Lexicon Building Process for Machine Translation between 'New' Languages," in *Proc.* 5th Conference of t*he Association of Machine Translation in the Americas Conference (AMTA)*, October 2002, London, UK, p 177-186

[5] L.R.H Chapmen, *English Grammar and Excercises*, 83rd ed, Longman group UK Ltd. London, 1994.

[6] InternetStudios (2012) English language history [Online]. Available: http://www.englishlanguageguide.com/english/facts/history/

[7] T.W. Wei-Jing Z. K. Papineni and S. Roukos, "Bleu: a method for automatic evaluation of machine translation," in Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL), July 2002, Philadelphia

[8] M.Prasad (2012), Unicode Converter - යුනිකෝඩ් පරිවර්තකය page [Online]. Available: http://unicode.malindaprasad.com/

[9] (2011) Google IME [Online]. Available: http://www.google.com/ime/transliteration/

[10] (2012) Unicode Sinhalese Editor and English to Sinhalese translator and converter to type in Sinhalese [Online]. Available: http://www.tamilcube.com/translate/sinhalese.aspx

[11] M.Kulatuga (2012) Madura Online Dictionary - Unicode Version [Online]. Available: http://www.maduraonline.com/

[12] A.M.Silva and Ruvan Weerasinghe, " Example Based Machine Translation for English-Sinhala Translations," in *Proc.* ICTer, 2008.

[13] B. Hettige, and A.S. Karunananda, "Computational Model of Grammar for English to Sinhala Machine Translation," in *Proc.* ICTer, 2011