

# Sinhala to English Language Translator

Dilshan De Silva\*, Asanga Alahakoon, Imesha Udayangani, Vishva Kumara, Devinda Kolonnage, Harindu Perera and Samantha Thelijjagoda<sup>++</sup>

Sri Lanka Institute of Information Technology

Malabe, Sri Lanka

\*Email: dindraraj@gmail.com

<sup>++</sup>Email: samantha.t@slit.lk

**Abstract**—This paper describes a machine translation system that is capable of translating a grammatically correct Sinhala sentence in to its corresponding English sentence. This is the first Sinhala to English machine translation system, which comes with features such as an inbuilt keyboard, an inbuilt dictionary, an integrated word processor based on Unicode fonts, a grammar tool, a Sinhalese grammar checker, an add word tool, and a debugging tool. With the expansion of the world, English has become an important language that people should learn, as the majority of the worldwide population understand and carry out their day-to-day work in English. In addressing this need, we thought of taking up the challenge of building, a Sinhala to English language translator. To build this system, we used the Transfer-based machine translation approach, which is a Rule-based approach. At present, the system has achieved a success rate of 75% with a corpus of 150 sentences.

**Keywords**—English, Sinhala language, transfer based machine translation, translator

## I. INTRODUCTION

Sinhala is a language spoken by nearly sixteen million people, amongst whom nearly thirteen million are native speakers. It has originated from one of the branches of the Indo-Aryan group of languages. It traces its origin to Sanskrit.

English, on the other hand is a West German language originating from England, used extensively as a second language and as an official language throughout the world. It has little co-existence with the Sinhalese Language, although it belongs to the Indo-European language family, it has much less in common with the Sinhalese language. A working knowledge of English has become a requirement in a number of fields, occupations and professions such as medicine, chemistry and consequently over a billion people speak English at least at a basic level.

Electronic forms of Sinhalese became available with the introduction of word processing software less than two decades ago. In contrast, the e-readiness of the English Language lies at the very top, being the language that has been studied most.

Although many Sri Lankans possess skills and sufficient knowledge to work in the Sinhala language because of their lack of knowledge in the English language they are unable to express their ideas and thoughts in a globally recognized language such as English. Difficulties have often been faced by many at job interviews, appointments of personnel for

various positions and facing examinations in English etc. In addition, foreigners also face difficulties when trying to understand and communicate in Sinhalese when they visit Sri Lanka. It is to overcome such barriers that we thought of this language translator.

It is a far more difficult task to perform a translation of non-similar languages than similar languages. For example, translation from English to French or German is easier than translation from Sinhalese to English, as English, French and German are closely related languages, and are based on a similar alphabet.

This paper introduces a pilot machine translation system that translates grammatically correct Sinhalese sentences into grammatically correct English sentences in a consistent, quick, accurate and precise manner along with several other features to facilitate the users.

Language translation has received much attention recently because it enhances the user's capability of understanding electronic or various publications by other means in different languages. A considerable number of Machine Translation projects have been carried out in Sri Lanka over the past decade. While most have attempted to translate from English to Sinhalese, translations from Sinhalese to English have received less attention, with no successful systems having been built until now, although there are existing systems that provide dictionaries for single words.

The research focused on the Transfer-based machine translation approach, based on the idea of Interlingua, which is currently one of the widely used methods of machine translations. It is a Rule-based approach. By using this translation strategy, it is possible to obtain high quality translations, with accuracy within the region of about 90% (although this is highly dependent on the languages being considered).

When developing the system we have considered all three "persons" and all the tenses in the Sinhala language. There are thirty-two grammatical regulations in the Sinhalese Language to construct a grammatically correct Sinhalese sentence(s), out of which twenty are for active voice, and twelve for passive voice. Shown below are some example sentences from the active voice, which cover the grammatical rules of Sinhalese:

මම දෙදර ගියෙමි.

නුඹ සාසලට යන්නෙහි.

මිනිස්සු වැඩ කරති.

Out of the twenty regulations for the active voice, this system handles sixteen of them. The balance is in the process of consideration. It has also been designed to handle sentences with adjectives, adverbs and nouns inclusive of days, months etc.

By using our system, Sinhala to English Language Translator (SELT) users can easily obtain the meaning of a particular word, sentence, or paragraph just by clicking a button on the graphical user interface (GUI). Therefore, they do not have to get help or refer to conventional dictionaries and waste their valuable time and that of others.

## II. METHODOLOGY

We used Object Oriented methodology to design the system. We carried out the work in the Visual Studio.NET environment with Visual Basic .Net as the programming language and Microsoft Access 2003 as the associated database management system. The choice of technology helped in the rapid prototyping and experimenting required for this research. An additional advantage of using .NET is the ease of porting this to a web service, or building interfaces or additional components through other .NET compatible languages (such as C#).

As the first step in the translation process, the translation engine takes in the input text that may be a sentence or paragraph. Each sentence should be grammatically correct and separated with a full stop (.) or a question mark (?). This is essential for the translation engine to produce accurate sentences.

The translation engine then analyses the input text. Once the system notices a full stop or a question mark, it considers the text up to that full stop or question mark, as one sentence, and each word of that sentence is added to a sentence object, which is then added to an ArrayList. This procedure continues until all the sentences are included in to the ArrayList. If there is a space, comma, colon, semicolon, end of a double quote, or single quote then the translation engine identifies it as the end of a particular word.

“Fig. 1” illustrates the steps carried out by the translation engine when converting the sentence සිදු වූයේ වේගයෙන් පාසලට දුවයි from Sinhalese to English.

Before adding a word to the data table (G in “Fig. 1”) the translation engine first checks whether the word is in English or Sinhala. If it is in English then the translation engine considers that word as an animate noun and adds it to the Data table. If it is in Sinhalese, the translation engine first converts it into “Singlish” (C in “Fig. 1”). Then it will be stemmed (a process of removing one character at a time from the end of the Singlish word) until it is found in the database. When it finds that word, it will add the information of that word such as whether it is a noun/verb, animate/inanimate, person, sex, etc to the Data table. Eventually for each sentence there will be one table, in which each row will contain details of a particular word.

The translation engine then joins words (H in “Fig. 1”) as necessary – for example joining an adjective with the noun

after it, or joining an adverb with the word after it. When joining certain types of words there will be no interchanging of the two words – for example, when displaying the output in the case of an adjective. In other cases, for example when an adverb is present, there will be an interchanging of the two words to comply with English grammar rules. Likewise the translation engine joins any words, other than nouns or verbs, such as prepositions, ownership nouns, definite plurals etc accordingly. The postfix of the final word – i.e. – the final letters of the Singlish word that are not included in the database – is extracted and used to determine the context of the sentence (I in “Fig. 1”). After that, if there are any objects in the sentence they will be formatted accordingly. eg: if the object is gedaraTA, in the output “to” will be added in front of home because the Sinhala word ends with “TA” likewise when the Sinhala word ends with “nTA”, “uTA”, “gee”, “ee”, “ehi”, “ka”, “hi”, “sin”, “kin”, “gen”, “en”, “n”, “nak”, “k”, “aava”, “uva”, “nva”, “vala”, “aa” the object will be formatted accordingly (J in “Fig. 1”). Table I illustrates different English words that are used when translating the above postfixes of Sinhala words.

TABLE I. ENGLISH WORDS FOR POSTFIXES OF SINHALA WORDS

Sinhalese	Singlish	English
ක්	k	a / an
කින්	kin	by a / by an
නක්	nak	on a / on an
උව්	uva	a / an
එන්	en	from a / from an
සින්	sin	from
න්	n	from
ගෙන්	gen	from the
වල	vala	in
එහි	ehi	in a
ක	ka	in a
ඒ	ee	on / on the
ආ	aa	on
හි	hi	on
ට	TA	to
ගේ	gee	of / of the
ආවා	aava	the
එන්ව්	nva	the
න්ට	nTA	the
උට	uTA	a/an

The translation engine then makes the string pattern for the Sinhalese sentence, which it will then use to make the English string pattern of the sentence (k in “Fig. 1”).

The subject of the sentence is formatted next (L in “Fig. 1”) and will be added to a string. Next, the verb will be formatted according to the tense and the voice (active or passive) and once it is formatted it will be included in the same string that the subject was placed, according to the sentence pattern (M in “Fig. 1”).

Finally, if there are any more words other than the subject and verbs those will be put in to the same above mentioned string according to the sentence pattern (N in “Fig. 1”).

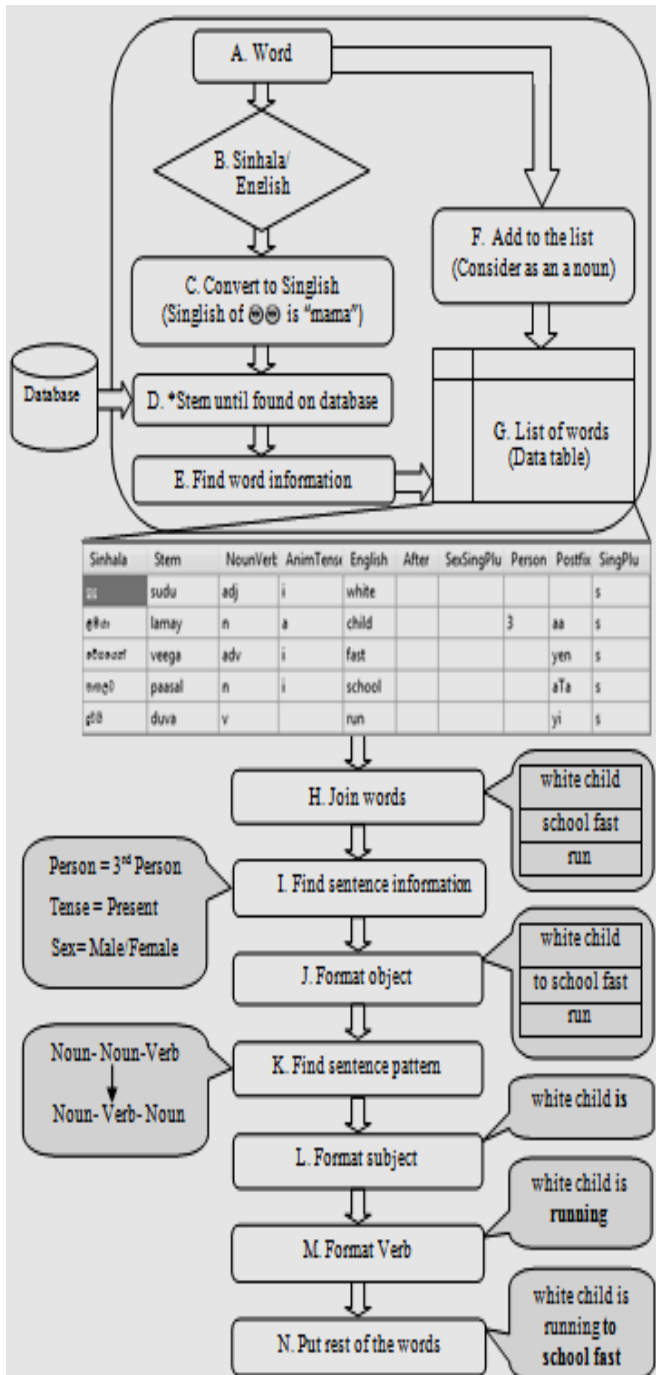


Figure 1. Translating the sentence  
සුදු ළමයා වේගයෙන් පාසලට දුවයි.

### III. THE SYSTEM AND ITS FEATURES

The main interface of the system is shown by “Fig. 2”. Apart from the translation process, this system contains features such as:

a *debugging tool* to assist future developers. “Fig. 3” illustrates how the Sinhalese sentence සුදු ළමයා වේගයෙන් පාසලට දුවයි is processed in the debugging tool.

an *integrated word processor* which provides an easy way of typing Sinhalese text, which can also translate grammatically correct Sinhalese sentences into English within a document itself and which provides a way of including the text, tables or images anywhere in the document. “Fig. 4” shows the module structure of the word processor.

an *in-built keyboard*, which provides an easy way of typing the Sinhalese sentences in English where the Sinhalese characters will be generated automatically. “Fig. 5” illustrates the process of the keyboard.

a *grammar checker* to verify the grammar of an input Sinhalese sentence or paragraph(s). “Fig. 6” illustrates the process of the grammar checker.

a *grammar tool* is also available which helps the user to find information about both Sinhala and English sentences – for example whether a particular word is a noun, verb, adjective - and provides all the materials for a user to do self studying of Sinhala and English languages. “Fig. 7” illustrates the process of the grammar tool.

an *in-built dictionary* has been included, which gives similar words to the word that the user entered, in addition to providing meanings for both English and Sinhala words, and example sentences for each word. Apart from these, it also provides a way of pronouncing an entered English word. “Fig. 8” illustrates the process of the dictionary.

an *“add-word” tool* which provides the user an easy way of adding words to the vocabulary through the front end.

We have also utilized Unicode Sinhalese fonts in order to facilitate the copying or transferring of the original and translated text to other applications.

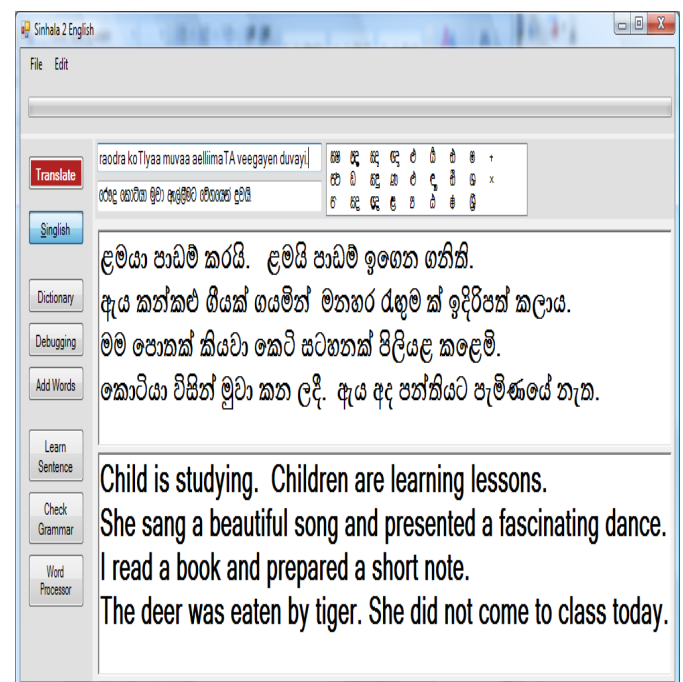


Figure 2. Main Interface of the Translator

Debugging									
Sentence	Sinhala	Stem	Noun/Verb	Anim/Tense	English	Age	Sex	Person	Sing/Plu
සුදු ළමයා	sudu lamaya	n	a	white child			3	aa	s
වේගයෙන් පාසලට	veega passala	n	i	to school fast				aTa	s
ඳුවයි	duva	v		run				yi	s
වෘත්තය				Present			Male	3	s

Figure 3. Debugging tool

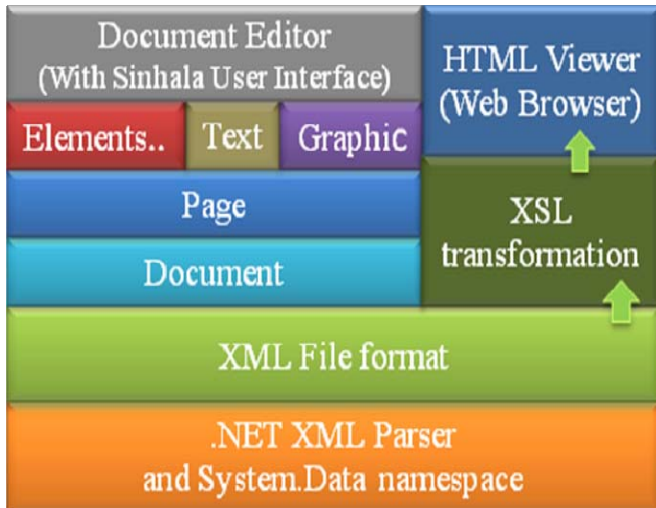


Figure 4. Module structure of the word processor

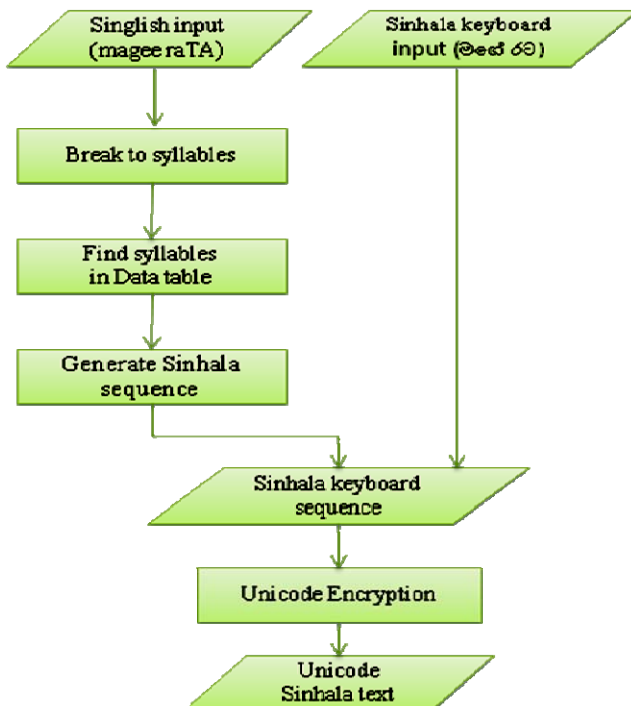


Figure 5. Process of the keyboard

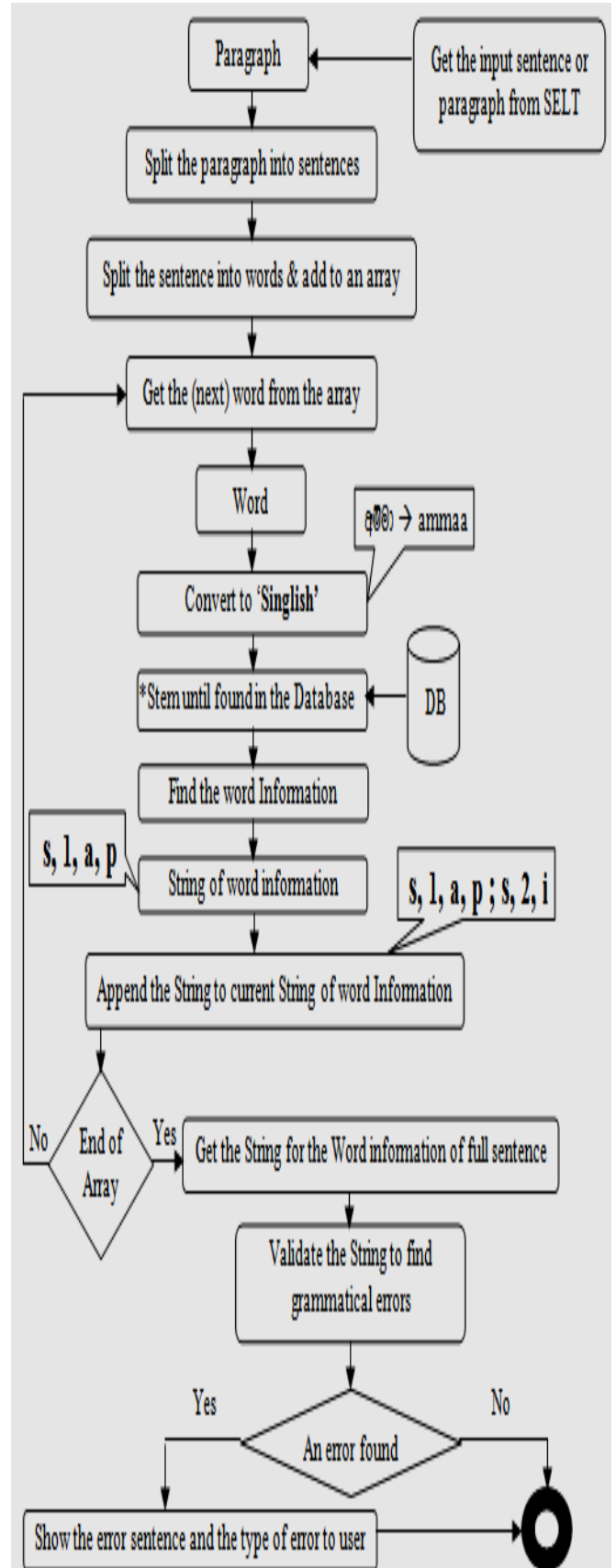


Figure 6. Process of the grammar checker



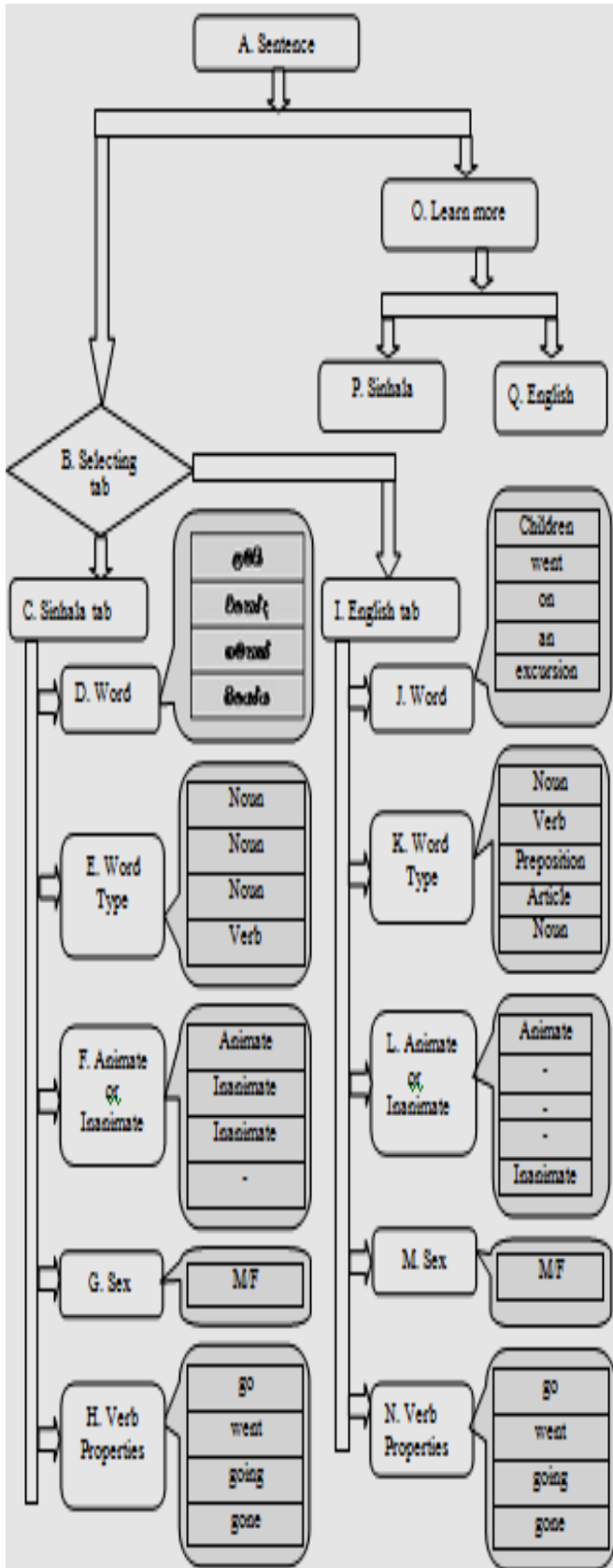


Figure 7. Process of grammar tool for  
 ළමයි විනෝද ගමනක් ගියේය sentence.

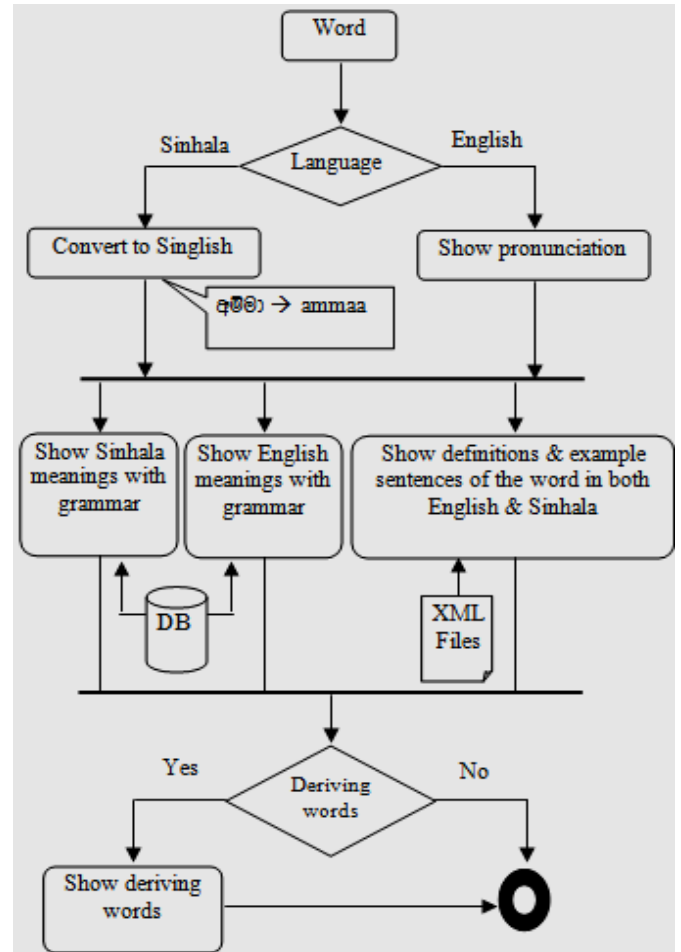


Figure 8. Process of the dictionary

#### IV. RESULTS AND DISCUSSION

##### A. Success

Current systems are incapable of producing output of the same quality as a human translator (particularly where the text being translated uses casual language) because a machine does not have the human dimension required for dealing accurately with the subtle nuances of natural language. However, a translation test for 150 grammatically well-balanced sentences was conducted using this system, and an acceptable translation result with 75 percent was achieved.

Table II contains some example sentence patterns that were correctly translated by the system. In the first sentence the female subject “ඇය” does two actions simultaneously. Since there are two verbs in this sentence, it will initially be divided into two parts and in the output those two parts will be joined with the “and” connector along with their English meanings. In the second sentence because of the conjunction “ඔ” initially the sentence will be separated into two parts. Then the English meanings of these two parts will be joined with the English meaning of the conjunction. When joining, the parts will be interchanged. The next set of sentences are examples of a sentence that is in passive voice, a sentence that is in negative form, a sentence with an ownership noun, a

sentence that has more than one noun connected with “ද”, and a sentence with a group word, respectively.

TABLE II. SOME EXAMPLES FOR CORRECTLY TRANSLATED SENTENCES

1. ඇය කන්කලු ගීයක් ගයමින් මනහර ධනුමක් ඉදිරිපත් කළාය. She sang a beautiful song and presented a fascinating dance.
2. ඔහු නිතර පාසලට පැමිණි බව තුරුවරයා දනී. The teacher knew that he came to school always.
3. නරියා වීසින් තාවා කන ලදී. Rabbit was eaten by fox.
4. මා අද පාසලට ගියේ නැත. I did not go to school today.
5. ඇය කුරුල්ලන්ගේ හඬට ඇඟෑම්කන්දී සිටියාය. She listened to bird's sound.
6. අම්මාද බුඹද මමද පන්සලට ගියෙමු. Mother, you and I went to temple.
7. ගුවන්යානා පෙළ අහසේ පියාසර කරයි. Line of aeroplanes are flying on the sky.

B. Future Research Directions

The system cannot get accurate translations for sentences with verbs such as “ඇත”, “නිමේ” and “සිටී”, wishes (second sentence in Table III), proverbs (third sentence in Table III), question form sentences (fourth sentence in table 3), orders such as the fifth sentence in Table III. These are the areas that we hope to look into in the future. We have already begun work in some of them. In addition, more focus has to be placed on the passive voice, as the current translator is mainly focused on the active voice sentences in the Sinhala language. It will be worth mentioning that we have progressed this far within a short period - about nine months.

TABLE III. EXAMPLE FOR INACCURATELY TRANSLATED SENTENCES

1.ලංකාව ඉන්දියන් සාගරයෙන් වටවී ඇත.
2.ඔබට දීර්ඝායුෂ ලැබේවා!
3.පුහුල් හොරා කරෙත් දැනේ.
4.පරිගණකයට මිනිස් සංස්කෘතිය වෙනස් කළ හැකිද?
5.පාරේ දකුණු පැත්තෙන් යනු.

The probability of getting an accurate result from this system depends on the number of sentence patterns that it contains, in other words, the more patterns, the more accurate the output. By increasing the number of words in the dictionary, users will have the benefit of being able to get more meanings and example sentences in Sinhalese and English. The speed of the grammar checker tool can also be optimized. The grammar tool can be further developed in a way that it would enable a foreigner to study Sinhalese grammar in the English medium.

V. CONCLUSION

This system helps to remove the barrier between English and Sinhala languages. It proposes a way of how technology based learning can enhance the quality of the current learning system for students who seek to understand the English language.

This Sinhala to English translator will help users to convert the Sinhala words, sentences or paragraphs to English language easily and to facilitate their learning.

By using structured and well-designed algorithms and necessary computer instructions to analyze, parse source sentences and perform translation, the Sinhala to English Language Translator can provide the above mentioned benefits to its users.

ACKNOWLEDGMENT

We would like to thank Dr. Samantha Thelijagoda, our project supervisor and Mr. Jayantha Lal Amararachchi the lecturer in-charge of the subject Comprehensive Design and Analysis of the project unit, for their enormous support while working on this project.

We would be failing in our duty if we do not mention Mrs. Chandra Perera, a teacher at Vishaka Vidylaya and Mrs. Ranjani Senevirathne who were of great assistance, with regard to guiding us to the correct form of the Sinhala Language.

Last, but not least, we are deeply indebted 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] S. Thelijagoda, Y. Imai and T. Ikeda, “Japanese-Sinhalese machine translation system Jaw/Sinhalese,” Journal of National Science Foundation Sri Lanka, vol.35, no.2, June 2007.

[2] N. Samarasinghe, English for Public Services, 8th ed., Department of Official languages, 2006.

[3] W. S. Karunatilake, An Introduction to Spoken Sinhalese, M.D. Gunasena and Company Ltd, Colombo 11, Sri Lanka, 1990.

[4] R. Murphy, Murphy’s English Grammar, 3rd ed., Cambridge University Press, The Edinburgh Building, Cambridge CB2, UK, 2004.

[5] S. O. Fernando, Sammana, Wara nonamena Nipatha, January 1994.

[6] S. O. Fernando, Sammana, Kriya pada igena ganimu, January 1994.

[7] S. O. Fernando, Sammana, Sinhala Nouns year 11, March 1994.

[8] D. Dissanayaka, Rivira, Adverb, October 2007.

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

[10] Machine Translation. (2008, October 19). [Online]. Available: [http://en.wikipedia.org/wiki/Machine\\_Translation](http://en.wikipedia.org/wiki/Machine_Translation)

[11] English Language. (2008, October 19). [Online]. Available: [http://en.wikipedia.org/wiki/English\\_language](http://en.wikipedia.org/wiki/English_language)

[12] English Grammar. (2008, October 19). [Online]. Available: [http://en.wikipedia.org/wiki/English\\_grammar](http://en.wikipedia.org/wiki/English_grammar)