GLOBAL ACADEMIC RESEARCH INSTITUTE

COLOMBO, SRI LANKA

GARI International Journal of Multidisciplinary Research

ISSN 2659-2193

Volume: 03 | Issue: 02

On 30th June 2017

http://www.research.lk

Author: Lahiru Abeyrathne, Surangi Edirisinghe, Rumesh Premachandra,
Apsaari Warsha, Nalaka De Silva, S. Thelijjagoda
Faculty of Computing, SLIIT, Sri Lanka

GARI Publisher | Machine Learning | Volume: 03 | Issue: 02
Article ID: IN/GARI/ICET/2017/140 | Pages: 70-78 (09)
ISSN 2659-2193 | Edit: GARI Editorial Team
Received: 24.03.2017 | Publish: 30.06.2017

Spell and Grammar Checking Tool for Sinhalese
අකුරු සෝදුව - සියඛස සුඛසක් කරනු රිසියෙනි

Lahiru Abeyrathne[1], Surangi Edirisinghe[1], Rumesh Premachandra[1], Apsaari Warsha[1], Nalaka De Silva[2], S. Thelijjagoda[3]

Department of Information Technology[1], Department of Information Systems Engineering[3], Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka.
Department of Sinhala, Faculty of Arts[2], University of Colombo

***Abstract* – Sinhala, as a natural language is a complex and a rich language derived from Pali and Sanskrit. According to the historical linguistics Sinhalese belongs to the Indo-Aryan family. When the language comes to the usage there are two forms/ verities as written and spoken. According to the linguistics this phenomenon is called "diglossia".When considering about the written context, it is a complex language that comprises of many spelling and grammar rules where the correctness of the formal writing totally depends on these well-defined rules. A word phrase, a sentence or a paragraph will perceive different meanings according to the syntaxes and semantics used. Therefore it is very important to ensure the spelling and grammatical correctness to deliver the desired meaning to the audience. Due to this high complexity, it takes a considerable time to manually proof read the content of a written context. The requirement of an automated mechanism to perform this task has been emerged for the Sinhala language many years back, with the technological advancement. Well-developed spell checking applications are available for the languages like English Tamil and Chinese. But, due to the morphological richness of the language, the applications implemented to process the Sinhala language is in its infant stage and also there is no evidence for the existence of grammar checking applications for the Sinhala language. Major drawback of the existing applications developed for the spell checking functionality of the Sinhala language is, they lack with resources to explore all the misspelled words provided to the application. Although there are predefined spelling rules in the Sinhala language, it has been difficult to come up with a rule based solution for the Sinhala spell checking.**

Therefore this research is intended to implement a web based real time system to check the spelling and grammatical correctness of a context written in Sinhala language. The spell checking component will follow up a data driven approach to eliminate the difficulties faced in existing spell checking applications and follows a rule based lexicon analysis methodology to come up with a novel approach to check the grammatical correctness. This novel application will facilitate the end users to check the spelling and grammatical correctness of the inputs they provide to the system and it will also provide suggestions to correct the mistakes appearing in the input.

The spell checking component follows a three stepped implementation approach as data gathering, spell checking and at the last phase the output is subjected to several optimizations to enhance the result. Since the grammar checking component is following a rule based approach, a corpus has been implemented by collecting different sentence types. The lexicons has been implemented based on these sentence types. It has been identified 6 types of noun phrases, 11 types of verb phrases, 4 types of adjective phrases and 9 adverb phrases to develop the lexicon.

In order to compare the performance of implemented system with the existing spell checking applications, several test cases were used. Other experiments were performed to analyze the functionalities of the application after integrating with the grammar checking component. All the test cases were executed in several operating systems to ensure the compatibility of the application. Also it was founded that the time taken to process the input and to provide the output varies according to the usage of optimization techniques.

As it is planned to implement the system that it can be easily understandable and usable to the end users, the market value of the outcome will increase. The comprehensiveness, reliability and the novelty of the outcome is planned to be monitored throughout the project. Beyond that, our target audience has been identified as newspapers, government sectors including all the ministries departments and authorities, banking sector of Sri Lanka, Schoolchildren and teachers and all people who wishes to deal with Sinhala language. The ultimate goal of the research will be protecting the mother language to the betterment of the future generation.

Keywords— Natural Language, Sinhala, Lexicons, Natural language processing, Machine learning

I. INTRODUCTION

As Sinhala is the official and the national language of Sri Lanka, most of the people are using Sinhala language in their day-to-day activities. Often the government ministries, departments and authorities mainly rely on Sinhala language to perform their daily tasks such as letter

writing, preparing reports, communicating between employees etc. It is a complex language where the correctness of the formal writing totally depends on many grammar rules. A word phrase, a sentence or a paragraph will perceive different meanings according to the syntaxes and semantics used. [1] Because of that, it is very important to maintain the correctness of the language in order to preserve the intended meaning. Lack of knowledge about the Sinhala language has made people to spend more time on activities such as letter writing. Often this situation causes for ambiguities in the outcome.

The main intention of this project is to provide a web based real time system for the end users to check the correctness of the spelling and grammar of Sinhala language. Users will be able to type in or paste a word, word phrase or a paragraph in to the system to check for the correctness. Due to the real-time checking, it will reduce the time taken for making corrections of the written content. In addition, it is planned to provide real time suggestions for the end users when typing a sentence.

Since Sinhala is the national language of Sri Lanka, most of the activities at government authorities, ministries and departments are carried out in Sinhala. Majority of the people are using Sinhala for their daily routines. Most of the activities at government authorities, ministries and departments are carried out in Sinhala. Such as letter writing, preparing reports, communicating between employees etc. Sinhala language consist of many grammar rules and syntaxes in writing. Most of the people don't know how to use those grammar and syntaxes rules correctly. Therefore the expected meaning can be different due to the incorrectness of grammar. Because of that, it is very important to use the language correctly in order to preserve the intended meaning.

In Sinhala letter writing people are always typing what they want. But they don't know if the sentences are grammatically right or wrong. Typing whole thing, then people check the corrections. Therefore when manual checking, the time consumption is high and then it results in low effectiveness. If there is any tool or system to check those corrections in real time, then there will be no time waste and also it will be very effective.

The outcome of the project will be useful in order to increase the efficiency of clerical activities at the above-mentioned places. Beyond that, we have identified our target audience as newspapers, government sectors and banking sectors of Sri Lanka, especially the Central bank, school children, teachers and all individuals who wish to deal with Sinhala language. As Sinhala, language is the constitutionally recognized official language in Sri Lanka the expected outcome can be used in the real environment by the end users. The long time results of the outcome will lead towards protecting our mother language for the benefit of the future generation.

The field of study that focuses on the interactions between human language and computers is called Natural language Processing or NLP [2] in short. It sits at the intersection of computer science, artificial intelligence and computation. linguistics. NLP way for computer analyze, understand and A grammar checker is one of the basic Natural Language processing applications or tools for checking the syntax of language. The natural language processing field is relatively new in Asian language and lot of tools are yet to be developed. One of these is a grammar checker. In general, rule based and machine-learning techniques have been used for developing grammar checkers. Because the field of natural language processing for Asian languages is limited, along with the complex grammar, presents some problems when developing a grammar checking system for the Sinhala languages. Different techniques have been used for the development of these systems.

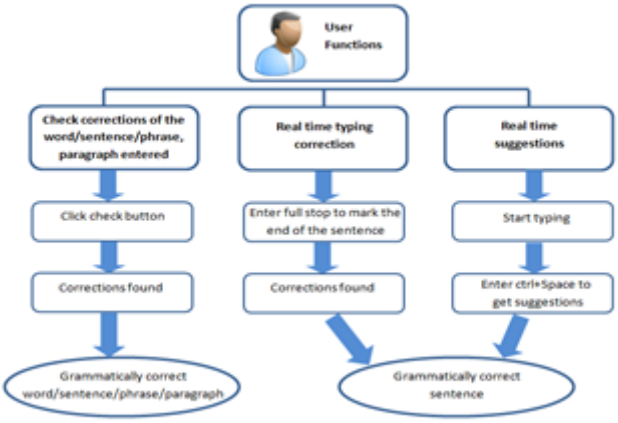
II. BACKGROUND STUDY

Sinhala is an Indo-Aryan language spoken by about 16 million Sinhalese people in Sri Lanka. It is also used as a second language by another 3 million people belonging to other ethnic groups in Sri Lanka, where it is one of the official and the national languages, along with Tamil. There are also considerable number of Sinhala speakers, writers in Singapore, Thailand, Canada and the United Arab Emirates. Sinhala language has its own writing system, which is an offspring of the Brahmi script. Maldives and Dhivehi are the closest relative languages to Sinhala. Further, Sinhala scripts are the world's 16th most creative alphabet among today's functional languages [3]. Committee on Adaptation of National Languages in Information Technology (CANLIT) arrived at identifying the Sinhala alphabet as having 16 vowels, 2 semi consonants and 41 consonants. Additionally 13 consonant modifiers were also identified. A new character to denote "fa" (ෆ) was introduced [4].

In any kind of language processing application, a lexicon plays a major role. Rather than using traditional approaches, a corpus-based approach has many benefits. According to the research paper that we studied on this approach the lexicon developed for Sinhala was based on a text obtained from a corpus of 10 million words [5].This lexicon has been implemented in XML according to the specification given in Lexical Mark-up Framework [6]. In this lexicon, sub category (subclass). Each Lemma has two feature attributes namely citation form and pronunciation. Word Form has several feature attributes called written Form, which is the orthographic representation, pronunciation, number, gender, person, definiteness and case of the particular word form.

Although there are grammar checking tools available for other languages such as English, the unavailability of a method to check the grammar of Sinhala language at real-time is still not prevalent. The complexity of the Sinhala grammar seemed to be the main reason for this. Since a corpus based lexicon for Sinhala language has been implemented, it can be used to drive meaning from human

language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognized, relationship extractions, sentiment analysis, grammar checking and topic segmentation follow a data driven statistic based approach. A specific method to configure the end of a sentence needs to be implemented in order to analyze an input by identifying each sentence. Additionally the system needs to be implemented in a way that it offers suggestions for the end users by identifying the pattern of the provided input. The currently available lexicon can be further enhanced by adding new words. According to the enhanced lexicon, the spell checking functionality should be implemented while making required updates to the lexicon.



Sinhala language contains lot of spelling and grammar rules. Therefore it is difficult to implement all these rules according to a specific order [7]. Since existing systems facilitate spell checking, this system provides spell checking together with grammar checking. The spell checking functionality has been implemented in a data driven approach in “සුබස” which is an existing system. There are other existing systems such as spellchecker.net and “මධුර”. Also we have studied a corpus based Sinhala lexicon in order to implement the initial step of spell checking. This has gained much more effectiveness than the traditional approaches. The words extracted from the corpus have been labeled with parts of speech categories defined according to a novel classification proposed for Sinhala. Project "මහරාවණා" done by University of Moratuwa has implemented simple grammar checking scenario for just simple three words sentence. Since it is a ruled based system, it is very difficult to make further improvements to that system due to the heavy scope of the Sinhala language. A proper lexical analyzer has not been implemented for Sinhala language at present. UCSC, UOM, SLIIT are engaged in interesting researches related to this area.

In order to start the implementation of the system, the Unicode system developed by University of Colombo School of Computing has given us a significant support. Another research carried out by the same institution is Sinhala lexicon. This paper presents the importance of revisiting traditional Sinhala grammar in order to keep abreast of current changes. The paper also proposes a reasonable classification for Sinhala words. Also the research paper, a data driven approach to checking and correcting spelling errors in Sinhala provides an approach which is described based on n-gram statistics and is relatively inexpensive to construct without deep linguistic knowledge. This approach is particularly useful as there are very few linguistic resources available for Sinhala at present.

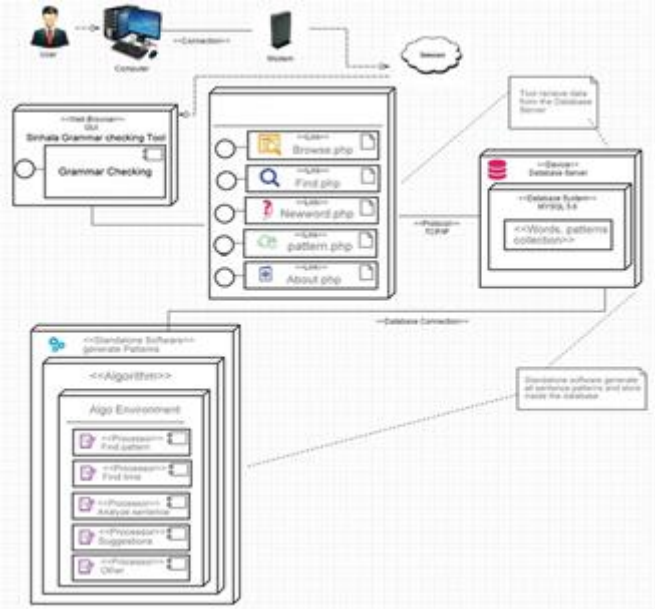
III. SYSTEM OVERVIEW

The proposed, “අකුරු සෝදුව - සියබස සුබසක් කරනු රිසියෙනි” (spell and grammar checking tool for Sinhalese) system provides a common web platform which can be used by anyone to achieve following practices in Sinhalese. The outcome will provide facility to end user to manually type or copy paste a word, word phrase or even a paragraph to check the to check the correctness of the spelling and grammar. System will provide the suggestion to the users to turn out the sentences.

This proposed system contains basically two parts.

- Spell checking
- Grammar checking

The output of the spell checking gain to complete the grammar checking part. User input always need to be only with Sinhala letters and punctuations. Numbers and any kind of symbols cannot be included. The necessary validations may implement for the input. System running on real time, therefore a user guide will be there. To increase system usability, efficiency, effectiveness and all the things required everything may include accordingly in the system.



By using a time to time updating dictionary, system will automatically identify the existing, non-existing and miss-spelled words. Then the system will give suggestions to each and every word accordingly, using the dictionary consisting of most of the spell rules. Therefore when writing, most of the spelling errors can be identified, specially "න, ණ" and "ල, ළ" rule. Then the system will predict the suggestions to the user accordingly and it's easy for user for ongoing. Introducing system algorithm always identified each sentence separately and checking grammar accordingly. Corresponding algorithms separately identify the SOV of the sentence.

After that it analyzes the sentence pattern for checking grammar correctness. In this section, system will use pre-defined grammar rules of lexicons. Lexicons will identify the sentence types separately. According to that, the system will predict suggestions to user for if they want further corrections. According to that make the suggestions to the user. Finally the system will produce correct spelled and grammatical sentence for user. Then user can ongoing with the system with provide user experience. It is more beneficial to "අකුරු සෝදුව - සියබස සුබසක් කරනු රිසියෙනි" team to develop this system further.

IV. SPELL CHECKING METHODOLOGY

There will be two main approaches to implement this task as "Rule-based" [8] and "Data driven". Due to the severe complexity of the Sinhala language, it is very difficult to define a fixed set of grammar rules. Well-developed spell checking applications are available for the languages like English Tamil and Chinese. But, due to the morphological richness of the language, the applications implemented to process the Sinhala language is in its infant stage. Major drawback of the existing applications developed for the spell checking functionality of the Sinhala language is, they lack with resources to explore all the misspelled words provided to the application. Although there are predefined spelling rules in the Sinhala language, it has been difficult to come up with a rule based solution for the Sinhala spell checking.

Therefore this research component describes a data driven approach to check the spelling correctness of the Sinhala text inputs provided by the users through a web based application in real time. This research component mainly follows an implementation approach based on probabilistic theorems and several other techniques has been applied for the optimization of the results.

The implementation process was carried out under two stages as:

- Data gathering
- Spell checking

During the data gathering stage, the data required to implement the corpora were collected. Spell checking stage of the implementation includes all the functions that were used to implement the core of this application. During the last stage of implementation, the REST API was implemented.

To make the application more accessible for the users and to make it easily integrated with the other research components of the group members. The spell checking functionality was implemented by identifying four major components of the approach.

- The language model
- The error model
- The candidate model and
- The selection mechanism

Within the system the suggestions will be listed under the drop down list named as "Suggestions". User will be able to select the required word from the drop down list to correct the spellings. User will be able to correct all the misspelled words by clicking on the button "Change All" or otherwise the user will be able to change a single word using the "Change" button. If the user is not satisfied with

the provided suggestions by the system, user will be able to ignore the suggestions by clicking on the "Ignore Word" or "Ignore all Words" button. Also if the suggestions are not available in the system, user will be facilitated to add the relevant new word to the system using the "Add Word" button.

V. GRAMMAR CHECKING METHODOLOGY

According to the Sinhala computational grammar rules the subject, verb and the object of the sentence along with the form of the sentence will be identified. Then in advanced grammar checking the active or passive voice of the sentence, masculinity or femininity of the sentence, singular or plural sentence will be identified. Then the system will check the correctness according to the identified scenario. The entire paragraph will be processes as described above accessing each sentence of the paragraph.

According to the Sinhala computational grammar [9] rules the subject, verb and the object of the sentence along with the form of the sentence will be identified. [10] Then in advanced grammar checking the active or passive voice of the sentence, masculinity or femininity of the sentence, singular or plural sentence will be identified. Then the system will check the correctness according to the identified scenario. The entire paragraph will be processes as described above accessing each sentence of the paragraph.

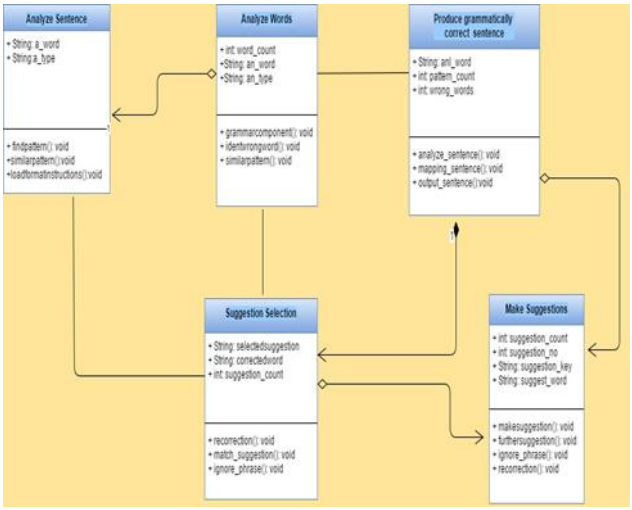
According to the grammar identified during the analysis phase by the lexicon [11], a new lexicon will be created to map the words to create a correct sentence. According to the updated lexicon the system will gradually develop to suggest different solutions to correct a grammatical error. Not only that, tried out another ways to do this work to analyze the effectiveness way of doing grammar correction. Using neural network system [12], tried out two different ways of implementations. Class based text classification and mathematical model text classification.

In class based text classification, sentences separated into the classes to train. In that case, need more data set with more classes. This text classification method already implemented for English [13]. It trained as greeting, good bye and sandwich kind of categories. Data set implemented as category wise. Then tried out to a new sentence, which matching with one of existing categories. Same thing applied to implement that on Sinhalese. There took "මම - mama" (I), "අපි - api" (We), "අම්මා - amma" (Mother), "ඔබ - oba" (You), kind of classes to train. Implemented sentences according to that classes to train. Then tried to identify new Sinhala sentence from a trained categories, if it is correct or not. Then that trained categories need to identify that new sentence. On that moment can identify that given new sentence in correct grammatical format or not.

In mathematical mode concept, develop Sinhala sentences with less number of Sinhala words. Give input as a Sinhala phrase and there allocated distinct unique numbers to the unique words. Then system will doesn't identify that Sinhala words or meaning, it only identify the word pattern with numbers of allocated for that word. Then trained correct Sinhala sentences as well as incorrect sentences. Given 1 as if it is a correct sentence, if that sentence in incorrect format system trained it as 0. Then expected result was kind of a percentage value. There user can provide new sentences and if it is a correct sentence output should be 100%. If it is an incorrect sentence output should be 0%. But this is a mathematical model, therefore system converges into that two ends with inter median value in between 0-100. If it is a correct sentence, at least that output percentage value should be beyond more than 50%. If it is an incorrect sentence output percentage value should be less than 50%. With considering about values of system output, then accuracy of that model also can be gain.

As the final step the system will virtually pronounce the corrected word, phrase or the paragraph. From this approach the users will get a clear understanding about the pronunciation also. Existing systems can only give facility to check spellings. Any Sinhala grammar checking tool or a system does not exist at present. The suggested system can go beyond that existing spell checking system with much more accuracy. The proposed system can give grammar checking facility with the automate suggestions for correct sentences to make this product valuable.

VI. SYSTEM IMPLEMENTATION



Analyzing grammar components to decide sentence pattern. System check if there is any matching patterns. If there are no matching patterns, system will choose similar types of sentence patterns. According to sentence pattern,

the system checks whether all grammar components are in its correct format. If not, the system will identify wrong words to offer suggestions. If there are any similar sentence patterns, the system check those scenarios as well. This class (produce grammatically correct sentence) get the input of the Analyze Words. By identifying sentence pattern and the grammar component words, class functionalities produce the main final outcome of project. This class (suggestion selection) get input from produce grammatically correct sentence class. In this class functionalities produce similar type of sentence suggestions to the user.

The grammar checker is one of the research project. This project includes three major applications in different ways. They are spell checking, grammatical checking and suggestions. A few researchers have looked at this type of project and reached a few of the system. These applications only check spelling, suggestion, and some relevant documents with their system. But in our application, we provide spell checking, suggestions and also Sinhalese rules checking. There is no grammar proofing tool available for the Sinhalese language. But some research and books are written about Sinhalese rules. But we cannot quickly check Sinhalese rules because many of the rules of Sinhalese rules are relevant in the book versions.

Accuracy is a widely accepted measurement to measure the quality of the software. The accuracy of the system depends on the input to the system as well as the output of the system. So the output of the system should be improved because usually the input cannot be controlled as it is entered by a third party user. So to do this task, a set of computer instructions and algorithms should be used to analyze the sentence or phrase entered by the user and produce the output within a short period of time to make the system more efficient. Also check the validity of the input and make suggestions if there are any errors.

In order to increase the performance and the speed, the system should be fast and cheap regardless of the number of sentences or phrases entered as the input. If the system does not have a good database design then it will lead to slower data retrieval. Then it will reduce the performance of the system. Therefore it is very important to have a powerful and a reliable database.

The "අකුරු මෝදුව" spell and grammar checker is used by people who have a good computer literacy as well as who does not have a good computer knowledge. So a person who does not have a proper IT knowledge should be able to use the system without any problem. In order to achieve this target the interfaces should be designed in a simple and a user friendly manner. Also a user guide is provided to use the software with ease and how to install it. Developing the system which supports fast installation will be easy for the third party. Grammar checking tool provide smooth and simple operations. In some error will occur in the particular section the system will generate the error message and inform to the user. In system development reliability plays a major role. For a high performance of a system reliability is must. At the same time multiple users can use the application.

In order to provide more efficient response to user requests the data base must be update. System data base needs update with new words and rules. User can add new words to the system. Without maintaining system the system team cannot achieve system goals. It is helpful to make the software substantial saving in the development costs and also to make the end product that is relatively low in complexity. The system is made so as to be very accurate in the tasks it is to perform. Correct Sinhala grammar given to the system.

VII. DISCUSSION

It has been found that this probabilistic approach has been able to provide more accurate results than the existing spell checking applications due to the used optimization techniques. Therefore the optimization techniques and the data driven probabilistic approach can be identified as the research findings that led the pathway to the success of this research component. The implemented spell checking functionality will mainly intended to provide the inputs to the grammar checking application which is the next phase of this research. When considered as an individual component, this functionality will help the users to ensure the spelling correctness of their documents during their day to day activities.

The grammar developed covers the default Sinhala sentence structure in the SOV order, and only they can be successfully parsed using the grammar developed. The rest of the sentence structures can't be parsed using the existing grammar. In natural language processing, dependency grammars are used to solve the free word-order problem.

Natural Language Processing (NLP) is an area of research that explores how computers can be used to understand and manipulate natural languages [14]. Developing a computational grammar for Sinhala can profit such efforts. Therefore in this research we report work carried out in developing a featurebased context-free grammar for Sinhala using the open source Natural Language Tool Kit, NLTK [15]. 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 people's namely first person (Uthtamapurusha) second

person (Maddamapurusha) and third person (prathamapurusha). Also there are nine cases in Sinhala such as Nominative (prathama), Accusative (karma), Instrumental (kaththru), Auxiliary (karana), Dative (sampadana), Ablative (avadhi), Genitive (Sambanda), Locative (adara) and Vocative (alapana)[16][17]. In written Sinhala there is no unique method for word segmentation. The linguistics literature reports on collections of rules for segmenting Sinhala words [18]. However most users of the language are not aware of these rules and do not follow them closely for word segmentation. For example the word-ending particle 'ය' is often used inconsistently. The Sinhala language has two types of verbs, namely shudda kriya 'pure verbs' and krudanta kriya 'participial verbs'. When a participial verb occurs in the sentence ending position there are two ways to write it. One is by separating the sentence-ending particle as in the case of 'ගියේය' "(he) went" and adding it to the participial verb as 'ගියේය'. Owing to this, it is desirable to have a word segmentation algorithm to check whether the text is in a normalized form before the CFG parser is employed.

There are number of sentence structures in Sinhala which do not contain a verb. These types of sentences end with adjectives, oblique nominal, locative predicates and adverbials among others, and the current grammar does not cover such non-verbal sentences of Sinhala.

In neural network class model, data set created as group wise (classes). Tried out separate classes wise data training. Then try to test that with new Sinhala sentence. Implementation as follows.

```
training_data = []
training_data.append({"class":"##", "sentence":"## ######## ##"})
training_data.append({"class":"##", "sentence":"## ### ###"})

training_data.append({"class":"##","sentence":"## ######"})
training_data.append({"class":"##","sentence":"## ######"})
print ("%s sentences in training data" % len(training_data))
print (training_data)
```

But the thing was it didn't work properly. As an individually identified that system need to identify the words of that Sinhala sentence also. In that case, if that data model contain lot of unique different Sinhala words system accuracy will going down. Therefore when preparing a data set need to implement that data set with some kind of a limitations with unique Sinhala words. Then system can identify Sinhala words with pattern. Then can hope higher accuracy than present observations.

Then system will output a percentage value of how much that sentence at that correct format. If it is more than 50%, then can think that sentence in accurately correct grammatical format. Two ends of 0% and the 100% is the results may offered grammatical wrong sentence and grammatical correct sentence respectively. Used Sinhala phrase in below.

Sinhala phrase: මම ගෙදර ගියෙමි ("mama gedara giyemi" – I went home). මම පන්සල් ගියෙමි ("mama pansal giyemi" – I went temple). මම පාඩම් කලෙමි("mama paadam kalemi" – I studied). මම ගෙදර ගියෙමු ("mama gedara giyemu"). මම පාඩම් කළෙමු ("mama padam kalemu"). අපි ගෙදර ගියෙමු ("api gedara giyemu" – We went home). අපි පන්සල් ගියෙමු ("api pansal giyemu" – We went temple). අපි පාඩම් කලෙමු ("api paadam kalemu" – We studied). අපි ගෙදර ගියෙමි ("api gedara giyemi"). අපි පන්සල් ගියෙමි ("api pansal giyemi"). අපි ගෙදර පන්සල් ("api gedara pansal"). මම ගෙදර පාඩම් කලෙමි ("mama gedara paadam kalemi" - I studied at home). මම ගියෙමි පන්සල් ("mama gedara pansal"). අපි ගෙදර පාඩම් කලෙමු ("api gedara paadam kalemu" – We studied at home). මම අපි ගියෙමු ("mama api giyemu"). මම ගියෙමි ("mama giyemi" – I went). අපි ගියෙමු ("api giyemu" – We went).

Some of sentences have proper meaning and some of them have' nt. Assign unique numerical values for Sinhala words and then trained the system.

Sinhala Word	Corresponding number of system identifying
මම ("*mama* "- I)	1
ගෙදර("*gedara* "- Home)	2
ගියෙමි("*giyemi* "- Went)	3
පාඩම්("*paadam* "- Study)	4
කලෙමි*("kalemi* "- Did)	5
ගියෙමු("*giyemu* "- Went)	6
කළෙමු("*kalemu* "- Did)	7
අපි("*api* "- We)	8
පන්සල්("*pansal* "- Temple)	9

Correct word format of Sinhala sentence trained as "1" and incorrect sentence format trained as "0". Sentence patterns and trained format as follows.

```
X = np.array([[1,2,3,0],[1,4,3,0],[1,5,6,0],[7,2,8,0],[1,5,8,0],[7,2,8,0],[7,4,8,0],[7,5,8,0],[7,2,3,0],[7,4,3,0],[7,2,
Y = np.array([[1],[0],[1],[0],[0],[1],[1],[1],[0],[0],[0],[1],[0],[1],[0],[1],[1]], dtype=float)
```

VIII. RESULTS

In order to compare the performance of implemented system with the existing spell checking applications, several test cases were used. Other experiments were performed to analyze the functionalities of the application after integrating with the grammar checking component. All the test cases were executed in several operating systems to ensure the compatibility of the application.

User entered incorrect spelled words = 50
System identified incorrect spelled words = 44
System not identified as incorrect spelled words = 6
Provide proper user suggestions = 36
Not provide proper user suggestions = 14

User entered correct spelled words = 50
System identified correct spelled words = 43
System not identified as correct spelled words = 7

Produce dataset as about around 150 sentences to train the system. Then those noun phrases, verb phrases, adjectives and adverbs and all of the words containing, separate into lexical format. Lexical developed by using that's words containing on data set. Then for all data given have been given as correct sentences in Sinhala language. In lexical analysis scenario gave following results.

Input correct sentences = 25, System identify only 2 sentences.
Input incorrect sentences = 30, System identify only 4 sentences.

Neural network mathematical model gave following results.

Provided correct grammatical order sentences – 25
System identified correct sentences – 8
Provided an incorrect grammatical order sentences – 25
System identified an incorrect sentences – 10

Therefore instead of a neural network mechanism, hope to go with machine learning techniques to do this. Then hope to go beyond with high accuracy level as well.

IX. CONCLUSION

This paper describes the development of a CFG for a non-trivial subset of Sinhala using the NLTK toolkit. Ten simple sentence structures were selected and used to design the grammar. Computational model of grammar for Sinhala language has been developed by considering the Morphology and the Syntax of the Sinhala language .Finite State Transducers (FST) and Context-free grammar (CFG) have been used to describe the computational grammar for Sinhala .The concept of Varanegeema (conjugation) is used as theoretical basics of the translation. In the future, it is hoped to use a morphological analyzer and a word segmentation algorithm to develop a more wide-coverage grammar for Sinhala.

According to lexical analysis methodology, there need to implement all of the Sinhala grammar rules for further going. Need to train more data set to system. Training method need less number of words with a huge number of Sinhala sentences. On that way system can understand that words separately. As well as grammar identifying logic also need to upgrade. Need well understand about Sinhala grammar rules to implement. Therefore as a team and individually prepared for learning these things. Lexical predefined grammar rules also need to recheck if that implemented way in exact correct way or not. Hope to go beyond 1000 Sinhala sentences to train the system. Expecting accuracy well going on above 60%. That well enough level for this research.

Neural network classes model also not in accurate. Also need to use less number of Sinhala words in sentence making. In that case cannot go beyond with some amount of a way. Because classes means there contain huge number of classes of in Sinhala. Therefore it is not in practical manner. It can construct into considerable amount of way and nothing else. That classes method supporting to English as much. But for the Sinhala Unicode doesn't work properly. Sometime those Sinhala words going to encrypt. Then system exit with the execution. This implementation also very difficult. Working scenario also cannot reliable. Efficiency and effectiveness at its lowest.

Neural network mathematical model also not provide accurate results as team expected. Percentage value not behave in expecting range of accuracy. That mathematical model not learning Sinhala language. That only leaning about numerical pattern of a correct and incorrect sentence. By allocating separate numbers for unique Sinhala words, cannot think about the ending point of this. Therefore implementation not in bounded. According to that system self-learning also not in bounded. Therefore cannot think about high accuracy in this method. Also it doesn't give much accuracy as expected. Instead of a Sinhala learning system, this mathematical model concept have some kind of a benefic. Because Sinhala word learning is very hard. But the thing is accuracy not much as expected. Therefore implementing scenario of this mathematical model need to implement on hard to get higher accuracy. On the other hand if the system not identifying Sinhala words and sentences, it is also not suitable for further going. System must need to clarify Sinhala words, meanings and sentence patterns. But are there any method to exceed that results accuracy hope to on that way also. Because neural networks are well performing against mathematical models.

REFERENCES

[1]A. Weerasinghe, "Language Technology Research Laboratory - UCSC",UCSC, 2010. [Online]. Available: http://ucsc.cmb.ac.lk/research-groups/language-technology-research-laboratory/.

[2]M. Kiser, "Infographic: What Are Algorithms? - Algorithmia Blog", Algorithmia Blog, 2016. [Online]. Available: https://blog.algorithmia.com/infographic-what-are-algorithms/.

[3]A. Wasala and K. Gamage, "Research Report on Phonetics and Phonology of Sinhala", Panl10n.net, 2007. [Online]. Available: http://www.panl10n.net/english/outputs/Working%20Papers/Sri%20Lanka/Microsoft% 20Word%20-%202_E_N_484.pdf. [Accessed: 18- Oct- 2017].

[4]V. Samaranayake, J. Dissanayake, A. Weerasinghe and H. Wijayawardhana, "An Introduction to UNICODE for Sinhala Characters", Nsrc.org,2003.[Online]. Available: https://nsrc.org/regions/ASIA/LK/03-Jan-2003_UCSC-Paper-on-Unicode.pdf.

[5]R. Jensen, J. Gatrell, J. Boulton and B. Harper, "Using Remote Sensing and Geographic Information Systems to Study Urban Quality of Life and Urban Forest Amenities", Ecologyandsociety.org, 2004. [Online]. Available: https://www.ecologyandsociety.org/vol9/iss5/art5/print.pdf.

[6]D. Naber, "A Rule-Based Style and Grammar Checker", Danielnaber.de,2003.[Online].Available:http://www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf.

[7]R. Weerasinghe, A. Wasala, D. Herath and V. Welgama, "NLP Applications of Sinhala: TTS & OCR", Aclweb.org, 2007. [Online]. Available: http://www.aclweb.org/anthology/I08-2142.

[8] Ladeforged, P., A Course In Phonetics, 3rd edn., Harcourt Brace Jovanovich College Publishers, 301, Commerce Street, Suite 3700, Fort Worth TX 76102 (1993)

[9]P. Taylor, A. Black and R. Caley, "The architecture of the festival speech synthesis system", Cs.cmu.edu, 1999. [Online]. Available: https://www.cs.cmu.edu/~awb/papers/ESCA98_arch.pdf.

[10]W. Karunathilake, "An Introduction to Spoken Sinhala - Professor W.S. Karunatillake", Scribd, 2011. [Online]. Available: https://www.scribd.com/doc/58658331/An-Introduction-to-Spoken-Sinhala-Professor-W-S-Karunatillake.

[11] Jayathilake, K., Nuthana Sinhala Vyakaranaye Mul Potha, Pradeepa Publications, 34/34, Lawyers' Office Complex, Colombo 12, (1991)

[12] T. Mikolov, M. Karafiat, L. Burget and S. Khudanpur, "Recurrent neural network based language model", Fit.vutbr.cz, 2010. [Online]. Available:http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf.

[13]F. Sebastiani, "Machine Learning in Automated Text Categorization", Nmis.isti.cnr.it,2005.[Online].Available:http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf.

[14]Sagar, B. M., Shobha, G., Ramakanth Kumar: "Context Free Grammar (CFG) Analysis for simple Kannada sentences" International Conference [ACCTA-2010] on Special Issue of IJCCT Vol. 1 Issue 2, 3, 4. (2010)

[15] Steven Bird, Ewan Klein, and Edward Loper "Natural Language Processing with Python" Analyzing Text with the Natural Language Toolkit. O'Reilly Media, 2009.

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

[17]R. Naccache, S. Anderson, S. Zhou, J. Taylor, M. DiGiacomo and T. Dickeson, "thesis.pdf | Morphology (Linguistics) | Parsing", Scribd, 2008. [Online].Available:https://www.scribd.com/document/322893774/thesis-pdf.

[18] Rajapaksha, D.: Sinhala bhashave pada bedima saha virama lakshana bhavithaya, Dharma Rajapaksha, (2008)