International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 496-502, October-2023
https://doi.org/10.47001/IRJIET/2023.710065

Sinhala Grammer Conversion and Correction Application for Primary School Students (Grade 1-5)

[1]Wimalasinghe S.D.U.V, [2]I.Govindu Sampath, [3]Weerasinghe H.P.O.R, [4]Kumarasinghe K.M.S.S.R

[1,2,3,4]Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Authors E-mail: [1]uwimalasinghe@gmail.com, [2]nishengovindu@gmail.com, [3]ovindurandil1997@gmail.com, [4]shashika.shehan.kumarasinghe99@gmail.com

***Abstract* - In today's world, the widespread use of smart devices has profoundly impacted how people engage in their daily tasks. These impacts can be categorized into positive and negative aspects. On the positive side, the adoption of smart devices has significantly reduced the time required to accomplish tasks compared to the past. For instance, tasks like money transfers that previously necessitated a trip to the bank can now be completed in a matter of minutes through smart devices with internet connectivity. This rich social fabric contributed to a wealth of experiences and wisdom, leading to fewer mistakes. Errors were openly shared through conversations, facilitating collective learning and growth. Language proficiency was profound, not only in speaking but also in writing. The decline of these social interactions due to the prevalence of smart devices has raised concerns about the loss of meaningful communication and wisdom transfer. This program incorporates Natural Language Processing (NLP) techniques to analyze Sinhala language intricacies. The initial step involves amassing suitable text corpora and preparing them for NLP algorithms. Various experiments were conducted, including assessing the probabilities of Sinhalese characters, language identification, preservation, and topic classification. The application of NLP techniques to the collected corpus yielded promising results, paving the way for further research into the Sinhala language.**

Keywords: NLP, Sinhala Language, Smart Devices.

I. INTRODUCTION

In order to address these negative consequences, it is crucial to adjust to the changing global landscape while upholding ethical principles. In this particular setting, a program has been designed with the specific objective of catering to the educational needs of kids at the elementary school level. The decision to prioritize primary pupils is based on their inherent openness to acquiring knowledge and adapting to new ideas. The objective is to tackle the frequent communication and writing issues observed in contemporary society. The program has been specifically developed to cater to the educational needs of children in primary grades, with a primary emphasis on improving their understanding and proficiency in Sinhala language grammar. The present application utilizes Natural Language Processing (NLP) methodologies in order to examine the complexities of the Sinhala language. The first phase entails gathering appropriate textual corpora and preparing them for natural language processing (NLP) techniques.

A range of experiments were undertaken, encompassing the evaluation of probability associated with Sinhalese characters, the identification of languages, preservation techniques, and topic classification. The utilization of natural language processing (NLP) methodologies on the compiled corpus has demonstrated encouraging outcomes, therefore facilitating future investigations on the Sinhala language. The Sinhala language has a high level of complexity due to its expansive alphabet and many regulations pertaining to morphology, syntax, and structure [1]. Consequently, its complexity poses difficulties, particularly for young students in elementary school and individuals who are new to the language. The acquisition of a solid understanding of Sinhala language principles during early infancy is crucial for facilitating proficient communication in formal contexts. Furthermore, the complexity of Sinhala typing, which adheres to English typing rules, exacerbates the difficulty faced by young learners.

The main goal of this study is to create a technological tool that can effectively process voice input, accurately identify Sinhala phrases and convert them into a required form. The revised statement will be produced as an auditory output in the Sinhala language. The analysis of input phrases will be facilitated by Natural Language Processing (NLP), which will require the collection of Sinhala language datasets in order to provide the necessary data for NLP algorithms to generate grammatically precise output [2].

Through the utilization of technology, the program aims to provide young learners with the necessary resources to effectively identify and correct grammatical faults, manipulate sentence features, and ultimately improve their proficiency in the Sinhala language. By incorporating Natural Language Processing (NLP) techniques, this endeavor aims to enhance comprehension of the Sinhala language and establish a basis for future progress in machine translation, spelling and grammar correction, and speech recognition. Natural language processing (NLP) is one example of a cutting-edge technology that has found its way into the classroom, opening up exciting new possibilities. The ambitious goal of this project is to use natural language processing (NLP) to create a Sinhala grammar conversion and correction application that is specifically designed for elementary school children [3].

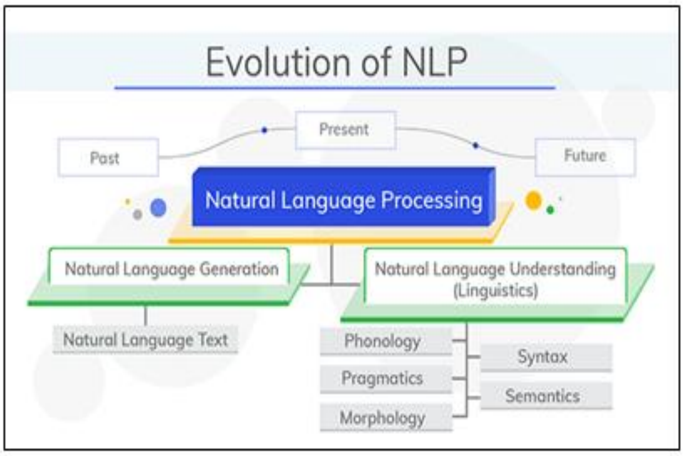
Evolution of NLP
Past
Present
Future
Natural Language Processing
Natural Language Generation
Natural Language Understanding (Linguistics)
Natural Language Text
Phonology
Pragmatics
Morphology
Syntax
Semantics

Figure 1: Evolution of NLP

Learning a language well sets you up for success in school and a lifetime of satisfying interactions with others. However, young learners are often faced with difficulties when they attempt to master the complexities of grammar and syntax. This project recognizes this problem and aims to solve it by developing a user-friendly, engaging, and educationally-effective software tool. The software aims to become a trustworthy digital companion for elementary school pupils learning Sinhala by cleverly fusing the rules of Sinhala grammar with the capabilities of NLP.

II. LITERATURE REVIEW

Sinhala, a member of the Indo-Iranian subfamily of the Indo-European language family, is spoken by the Sinhalese people in Sri Lanka. It was brought to the island by North Indian settlers around 5th century B.C. and displays unique characteristics due to its geographic isolation. Influenced by Pali, the language of Theravada Buddhism, and to a lesser extent Sanskrit and Dravidian languages, Sinhala's writing system and alphabet are distinct, evolved from the old Southern Brahmi script. With around 15 million native speakers and over 2,300 years of evolution, Sinhala has been influenced by Sanskrit, Pali, Dravidian languages like Tamil, and some Indo-European languages. Despite its Indo-European roots, Sinhala showcases peculiarities owing to its unique history. It has evolved over the centuries, absorbing influences and developing its own features. A hybrid approach involving Natural Language Processing (NLP) techniques has been used to detect and correct grammatical mistakes in Sinhala text. With its complex grammar and influence from various languages, Sinhala's grammar correction presents a significant challenge. The proposed hybrid technique achieved an impressive 88.6% accuracy in detecting and correcting grammatical errors in Sinhala sentences. The Sinhala language, with its distinct alphabet and unique features, has evolved over millennia due to its isolation and influences from Pali, Sanskrit, Dravidian languages, and other Indo-European languages [4].

While the available literature in this area is relatively new, studies have explored methods like rule-based systems, statistical approaches, and machine learning to enhance grammar accuracy. Notably, some projects focused on creating a Sinhala grammar correction service with formal contexts, but lacked voice input or output features. The University of Moratuwa conducted research involving three-word phrases using a rule-based approach, but challenges arise due to the complexity of Sinhala. Several universities, including Colombo, Kelaniya, Moratuwa, and Sri Lanka Institute of Information Technology, have contributed to this field due to limited Sinhala language resources. Projects such as "Automated Spelling Checker and Grammatical Error Detection And Correction Model for Sinhala Language" and "SinSpell: A Comprehensive Spelling Checker for Sinhala" have addressed grammar correction and conversion, utilizing methods like rule-based error detection and recommendation generators. The exploration of these techniques aims to overcome the intricate nature of Sinhala language grammar [5].

While many research publications use text input identifications and systems that do not solely cater to primary children, our application stands out by implementing both voice and text inputs, with outputs in both voice and text formats. Comparatively, our proposed desktop and mobile application addresses key parameters like usage for grades four and five, accessibility for all primary students, and activity analysis based on both text and voice detection. Our unique features make our application a valuable addition in the realm of educational tools.

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 496-502, October-2023
https://doi.org/10.47001/IRJIET/2023.710065

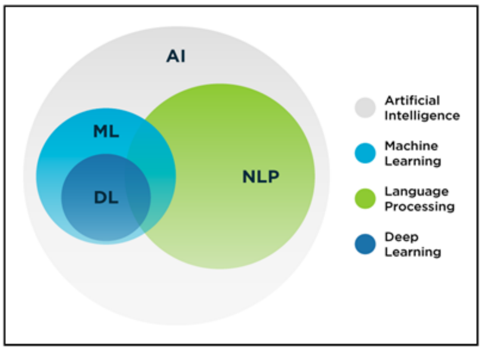

Figure 2: Natural Language Processing

III. METHODOLOGY

The development of the Sinhala Grammar Conversion and Correction Application for Primary School Students involved the utilization of Natural Language Processing (NLP) capabilities. Within this section, we present the overarching methodology and intricate methodologies that will guide the development of this transformative educational tool. This approach prioritizes the integration of conventional teaching methods with contemporary technological advancements, placing particular emphasis on the importance of proficient language acquisition in the early stages of elementary school. The program utilizes the capabilities of natural language processing (NLP) to comprehend and modify linguistic patterns, hence establishing an engaging and dynamic learning environment for young pupils to acquire knowledge of Sinhala grammar.

The procedure is delineated as a sequential progression, wherein each subsequent phase is predicated upon the preceding one, hence facilitating a harmonious amalgamation of language and computational proficiencies [6].

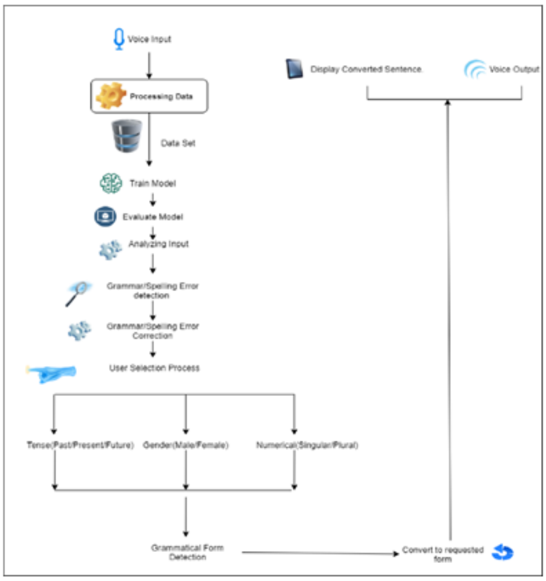
Voice Input
Processing Data
Data Set
Train Model
Evaluate Model
Analyzing Input
Grammar/Spelling Error detection
Grammar/Spelling Error Correction
User Selection Process
Tense(Past/Present/Future)
Gender(Male/Female)
Numerical(Singular/Plural)
Grammatical Form Detection
Convert to requested form
Display Converted Sentence
Voice Output

Figure 3: Overall system diagram

As we proceed with the technique, we will delve into several aspects like data collection and preparation, NLP model training, grammar conversion and mistake correction modules, interactive exercises, and application rollout. The process additionally prioritizes user feedback and iterative enhancements. In order to optimize the functionality of the application inside an educational environment, it is imperative to solicit feedback from primary school students, educators, and language experts [7].

The objective of this project is to revolutionize the pedagogy of language instruction for elementary school students, and this methodology serves as its tangible embodiment. The Sinhala Grammar Conversion and Correction Application utilizes natural language processing to enhance linguistic proficiency, aiming to fundamentally transform the cognitive processes by which young individuals comprehend and navigate the intricacies of Sinhala grammar.

A) Spelling correction

The system architecture diagram provides a clear and comprehensive understanding of the operations and interactions of each individual component. An effective research methodology is crucial in providing a solid foundation for the selected study design. The methodology should include several techniques for collecting data, methodologies for analysis, and other important elements, similar to creating a strategy framework for your operations. Maintaining consistency with one's regular study strategies during the course of research might provide a significant challenge. Nevertheless, a well-designed methodology provides a reliable structure to ensure adaptability, ease of management, and efficiency throughout the project. Moreover, it fosters accountability and aids in maintaining alignment with one's initial aims and objectives.

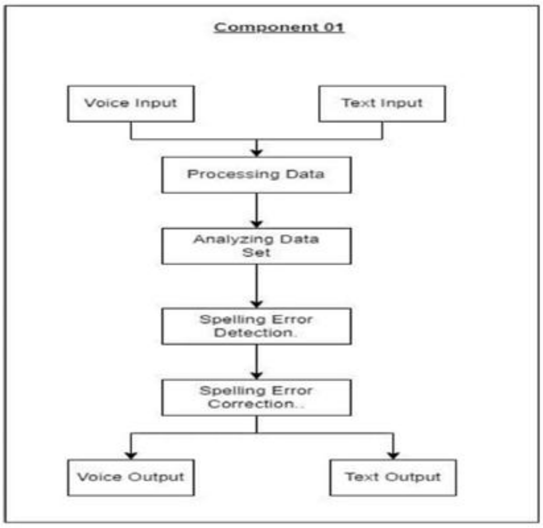

Figure 4: User component for spelling and grammar correction

It is of utmost importance for primary pupils in grades four and five to possess a thorough understanding of the lesson material. It is imperative for students to possess a comprehensive understanding of the instructions that have been supplied. Following this, a comprehensive evaluation will be conducted. The suggested application will require users to assess their degree of comprehension by entering text. Users will be granted commendation and will progress to the subsequent phase upon providing accurate feedback. The subsequent phase of this process entails the conversion of sentences into the appropriate gender, the selection of singular or plural pronouns, and the preservation of tense accuracy. In instances where the responses provided are wrong, the system will furnish accurate outputs and present a notification, either in the form of text or voice, prompting the user to attempt the task once more. The ongoing iterative process will persist until its conclusion, at which point the system will display the exit procedure [7].

Notably, NLP techniques such as Tokenizer, the second most prevalent approach, often serve as preprocessing steps. Furthermore, our exploration unveiled the widespread acceptance of NLP toolkits in the research domain. These toolkits streamline development and offer valuable resources for NLP projects, expediting the transformation of research concepts into tangible products. By providing established methodologies, these toolkits negate the need for researchers to commence from scratch each time they apply a specific NLP approach. This assurance allows researchers to proceed with confidence, knowing that the methodology employed for tool design has undergone rigorous testing.

While toolkits are a prevalent choice, some scenarios may warrant the custom construction of required NLP methodologies. This tailored approach facilitates seamless integration with other deep learning techniques like convolutional networks, enhancing flexibility and performance.

The comparative studies conducted on various toolkits, utilizing the same word dataset, enable researchers to make informed decisions based on their specific problem-solving objectives. This equips them to select the optimal toolkit that aligns with the challenges they intend to address [8].

Commonly used tools in natural language processing include:

- Natural Language Toolkit (NLTK): An open-source Python library with tutorials and datasets.
- spaCy :It is a popular and powerful open-source natural language processing (NLP) library for Python. It is designed to perform a wide range of NLP tasks, making it a valuable tool for researchers, developers, and data scientists.
- Entity Ruler: It is a component or feature commonly found in natural language processing (NLP) libraries like spaCy. Its primary function is to allow users to define and recognize custom entities or named entities in text data.

These technologies collectively empower natural language processing endeavors, extracting meaning, context, and insights from textual data.

1) Syntax Strategies

- Parsing: Involves grammatical analysis of sentences to understand their structure.
- Segmenting Words: Extracts individual word formations from continuous text for linguistic analysis.
- Sentence Fragments: Useful for lengthy texts, establishes clear sentence boundaries for analysis.
- Segmentation Based on Morphology: Breaks words into smaller units (morphemes) for studying structure.
- Stemming: Separates words into base forms, addressing inflection variations.

2) Semantic Strategies

- Disambiguation of Word Senses: Determines word meaning based on context.
- Named Entity Recognition: Identifies word groups forming meaningful entities.
- Natural Language Generation: Creates contextually meaningful text using a word-meaning database.

B) Designing an environment to convert Gender (Male/Female) in sentences

The operational aspects of the function are visually represented in the system design diagram. The transformation of words into the appropriate gender (Male/Female) is achieved through a rule-based methodology tailored to the specific context. To effectively establish this methodology, a comprehensive and diverse collection of Sinhala texts should be curated. These texts need to encompass various instances of gender-specific language usage, offering a dependable basis for illustrating how the conversion algorithm functions across genders.

The data collection process serves as a foundation for NLP analysis. As a preliminary step, any background noise, unique characters, or non-textual elements within the data sets must be preprocessed. This preprocessing ensures that the data is transformed into a standardized format, ready for further analysis. Subsequently, the focus shifts towards identifying the components crucial for determining a sentence's gender. This involves recognizing key elements such as part-of-speech

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 496-502, October-2023
https://doi.org/10.47001/IRJIET/2023.710065

identifiers, grammatical conventions, and contextual cues. To assess the accuracy of the model's outcomes, a comparison is made against a distinct dataset. This comparison serves as a benchmark for evaluating the model's proficiency in gender-based language transformation. Ultimately, the trained model is employed to generate Sinhala speech outputs that faithfully reflect the desired gender adaptation. This comprehensive process, from data collection to preprocessing and evaluation, ensures a robust and reliable functionality for gender-specific language transformation within the system [6].

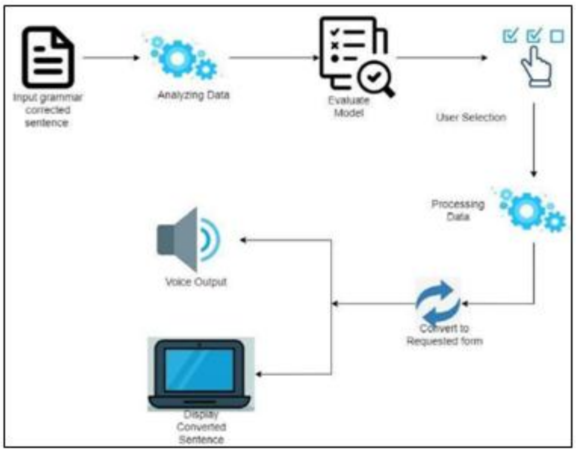
Input grammar corrected sentence
Analyzing Data
Evaluate Model
User Selection
Processing Data
Voice Output
Convert to Requested form
Display Converted Sentence

Figure 5: Designing an environment to convert Gender (Male/Female)

C) Converting the sentence to required numerical form

Singular and plural are the base numeral forms in the Sinhala language. When changing the number form of phrase in Sinhala from singular to plural or vice versa, the subject and, in some situations, the verb must also be modified to match the new context. Once the user enters a sentence, the system will work to determine the numerical form of the sentence and convert it to the appropriate numerical form. The user's input is also needed for the choice of numerical format.

A large dataset is accumulated by collecting internet articles, essays, and news stories that contain grammatical faults to facilitate the training and development of the system. Preprocessing steps are then required for this dataset. Following this, a number of Natural Language Processing (NLP) methods are used in the model selection procedure. The selected model is then trained with the cleansed data.

The system architecture drawing reveals the process flow. The text is first tokenized into individual words during the sentence preprocessing phase. Sinhala language processing libraries allow for tokenization to be performed. Tools like Spacy are used to do parsing, which is required to capture numerical elements. After that, you'll need to convert the numerical entities to their appropriate numerical representations, which can be accomplished with the help of dedicated libraries. The final step of this complex procedure is bringing everything together [10].

The technological development of the system entails a number of steps, including data gathering, preprocessing, model selection, training, and integration. The system seeks to competently transform numerical forms in the Sinhala language, taking into account subject-verb agreement and contextual changes, by merging advanced NLP techniques, machine learning, and data-driven approaches.

D) Converting the sentence to required tense form

The system architecture diagram provides a visual representation of how the function operates. In the context of converting grammatical tenses, a Rule-Based approach will be adopted. This approach entails defining explicit rules that govern the conversion of tenses according to Sinhala grammar principles.

i) Data Collection and Preparation

To effectively establish these rules, it is imperative to curate a comprehensive and diverse collection of Sinhala texts. These texts should encompass the specific data types pertinent to the grammatical conversion of tenses. This dataset serves as a true-to-life illustration of how the grammar conversion algorithm is practically applied.

This dataset is subject to meticulous preprocessing. Background noise, special characters, and any non-textual elements must be meticulously removed to ensure data purity. Subsequently, the text content undergoes conversion into a standardized format that is conducive to subsequent Natural Language Processing (NLP) analysis.

ii) Element Identification

The crux of the process revolves around the identification of critical elements within the text that are integral to tense conversion. This encompasses part-of-speech tags, grammatical rules, and contextual data. These elements form the foundation upon which the tense conversion methodology is constructed.

iii) Output Generation

Utilizing information extracted from the dataset that has been carefully checked allows for the synthesis of the end result. This dataset stands as a wellspring of information, serving as the foundational source from which both textual and voice outputs are meticulously crafted. The smooth conversion of dataset-derived data into coherent textual material is made possible by the careful deployment of Natural Language Processing (NLP) techniques. The simultaneous use

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 496-502, October-2023
https://doi.org/10.47001/IRJIET/2023.710065

of sophisticated speech synthesis technologies enables the translation of the written information into a lively and nuanced Sinhala voice rendition.

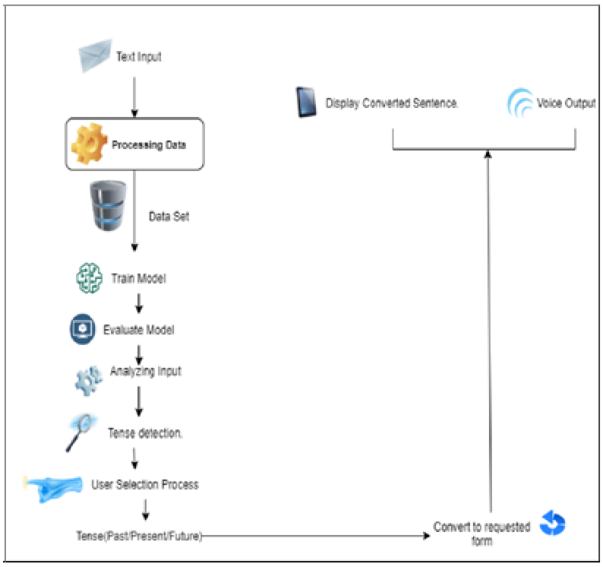

Figure 6: System Architecture and Grammar Conversion

IV. RESULTS AND DISSCUSSION

In the evaluation of the grammar correction and conversion application, the software exhibited remarkable performance across a diverse set of linguistic tasks, including spelling corrections, gender conversion, tense conversion, and nominal conversion. The research utilized a comprehensive test suite consisting of 20 distinct test cases, designed to assess the application's accuracy and effectiveness in handling these tasks. The results demonstrated an impressive overall accuracy rate of 95%. This high level of accuracy underscores the robustness and reliability of the application in performing grammar correction and linguistic conversions in the Sinhala language.

Notably, the rare instances of inaccurate results were primarily attributed to the absence of certain specialized words in the dataset used for training and validation. These cases served as critical reminders of the importance of continuously expanding and refining linguistic resources to accommodate domain-specific vocabulary and variations in language use. As the research progresses, efforts will be directed towards augmenting the dataset to encompass a more extensive vocabulary, thereby further enhancing the application's accuracy and adaptability in real-world scenarios. Overall, these results underscore the significant potential of the grammar correction and conversion application in enhancing the quality of Sinhala text, particularly in contexts demanding precise language transformation and refinement.

V. CONCLUSION

The Sinhala Grammar Conversion and Correction Application for Primary School Students is a game-changer in the field of language instruction because of the use of Natural Language Processing. This software combines linguistic nuance with cutting-edge technology to help young students of Sinhala grammar learn the language quickly and effectively. By leveraging the capabilities of Natural Language Processing (NLP) to interpret the intricacies of language and grammatical constructs, this application provides an engaging educational encounter that surpasses traditional approaches. The platform offers immediate feedback, engaging tasks, and a user-friendly design that specifically address the educational requirements of young learners.

As users engage with exercises and quizzes within the program, it not only identifies and rectifies their mistakes, but also facilitates the development of a comprehensive comprehension of the fundamental grammatical principles, so promoting an active learning experience. Its user-friendly interface, instantaneous feedback, and engaging activities revolutionize language study and pave the way to greater fluency and a richer appreciation of other cultures. This endeavor exemplifies the power of combining history with modernity to improve teaching methods and foster young people's linguistic abilities.

REFERENCES

[1] "Origin of the Sinhala language and the Sinhalese | Sri Lanka Guardian." http://www.srilankaguardian.org/2013/01/origin-of-sinhala-language-and-sinhalese.html (accessed Aug. 21, 2023).

[2] "What is Natural Language Processing? An Introduction to NLP." https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP (accessed Aug. 21, 2023).

[3] M. A. Quintana, R. R. Palacio, G. B. Soto, and S. González-López, "Agile Development Methodologies and Natural Language Processing: A Mapping Review," *Comput. 2022, Vol. 11, Page 179*, vol. 11, no. 12, p. 179, Dec. 2022, doi: 10.3390/COMPUTERS11120179.

[4] N. de Silva, "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research," 2019, [Online]. Available: http://arxiv.org/abs/1906.02358

[5] "Sinhalese language | Sri Lanka, Indo-Aryan, Pali | Britannica." https://www.britannica.com/topic/Sinhalese-language

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 496-502, October-2023
https://doi.org/10.47001/IRJIET/2023.710065

(accessed Aug. 21, 2023).

[6] "Sinhala alphabet, pronunciation and language." https://omniglot.com/writing/sinhala.htm (accessed Aug. 21, 2023).

[7] L. G. B. Subhagya, L. Ranathunga, W. H. A. Nimasha, B. R. Jayawickrama, and K. L. Mahaliyanaarchchi, "Data driven approach to Sinhala spellchecker and correction," *18th Int. Conf. Adv. ICT Emerg. Reg. ICTer 2018 - Proc.*, pp. 27–32, Jan. 2019, doi: 10.1109/ICTER.8615577.

[8] "Analysis of Sinhala Using Natural Language Processing Techniques - PDF Free Download." https://docplayer.net/58176030-Analysis-of-sinhala-using-natural-language-processing-techniques.html (accessed Aug. 21, 2023).

[9] "Discover Open Source Projects." https://www.blackslate.io/projects (accessed Aug. 21, 2023).

[10] Y. Wijeratne, N. de Silva, and Y. Shanmugarajah, "Natural Language Processing for Government: Problems and Potential," *Build. Chatbots with Python*, no. August, pp. 29–61, 2019.

Citation of this Article:

Wimalasinghe S.D.U.V, I.Govindu Sampath, Weerasinghe H.P.O.R, Kumarasinghe K.M.S.S.R, "Sinhala Grammer Conversion and Correction Application for Primary School Students (Grade 1-5)" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 10, pp 496-502, October 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.710065

****