Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

The Art and Science of Translating English to Singlish

D. I. De Silva, E. Weerasinghe, M. S. Shiraz, H. G. M. K. K. L. Karunasena, C. H. Zimmendra, O. A. Kumarasinghe

Dept. of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology Malabe, Sri Lanka

Abstract:- This paper describes the process of translating the mixture of Sinhala and English which is known as Singlish to grammatically accurate English which is done through the algorithm known as stemming. Furthermore, Transfer-based machine translation is the approach that's being used to deliver the result of Singlish to English translation. Singlish in the current generation is widely used throughout Sri Lanka considering that it is Sri Lankan's mother tongue, it makes sense that it's used with the combination of English. While translating Singlish to English there are also general features that tend to be very useful such as translating English to Sinhala, scanning text from a PDF document, a dictionary used to seek out words, and most importantly a feature that suggests the correct word while typing the words that need to be translated. The base technology that is being used is MERN which to elaborate includes Mongo DB, Express JS, React, and Node.

Keywords*: singlish, english, sinhala, translation, dictionary, stemming, scan.*

1. Introduction

As a Language that is mostly spoken only in Sri Lanka, Sinhala belongs to the Indic group of the Indo-Iranian subfamily of the Indo-European family of languages. Also described as an Indo-Aryan language and sometimes called Sinhalese, the same name as the people, it was brought to Sri Lanka by the north Indian peoples that settled on the island beginning in the fifth century B.C. Because of the relatively geographical isolation of these people from other Indo-Aryan tongues, Sinhala developed uniquely. It has been strongly influenced by Pali, the sacred language of Theravada Buddhism. To a lesser extent, it has also been influenced by Sanskrit. It also has borrowed words from the Dravidian languages of southern India, mostly Tamil. Sinhalese is written in its alphabet and script (abugida), which, like other South Indian writing systems, is derived from the ancient Southern Brahmi script [1].

English is a West Germanic language that originated from the Anglo-Frisian dialects and was brought to Britain by Germanic invaders (8th and 9th centuries AD). One second invasion took place by the Normans of the 11th century, who spoke Old Norman and developed an English form of this. That is why a large portion of the modern English vocabulary comes from the Anglo-Norman languages. A new vocabulary introduced at this time heavily influenced many organizations, including the church, the court system, and the government. European languages, including German, Dutch, Latin, and Ancient Greek influenced the English vocabulary during the Renaissance. The Old English period was from the mid-5th century to the mid-11th century, the Middle English period from the late 11th century to the late 15th century, the Early Modern English period from the late 15th century to the late 17th century, and the Modern English period from the late 17th century to the present [2].

Sinhala and English had their fair share of influence on each other's languages in the past due to the Colonial Period which Sri Lanka had to go through in the late 1800s to the mid-1900. After the independence, Urban areas and the rural areas of Sri Lanka saw a separation in the way that people speak. Colombo and the suburbs which are considered to be urban and Dambulla and Monaragala which are considered to be rural had their contrast in the way Sinhala as a language is spoken. In the Urban area, Sinhala is spoken with a mix of Sinhala words and a few English words. With the implosion of Technology and with the teen generation being fond of using technology to a greater extent [3]. Sinhala as a language saw a re-invention in Singlish (English and Sinhala Mixed)

Even though Singlish is spoken to a greater extent in the present days, the number of people who are speaking fluent English is on the lower side. Amidst the rich linguistic tapestry of Sri Lanka, it becomes increasingly evident that the acquisition and mastery of the English language hold substantial implications for the nation's economic prospects. Being considered as one of the best brains of South Asia due to the educational ecosystem which the country follows, the ability to communicate with different countries and different companies is efferently lower (due to the mother tongue-based educational system followed by the country). Due to this aspect, despite the country's low labor cost and high labor Functional knowledge, Sri Lanka may be unable to separate itself from the rest of the globe as a lucrative place for commercial and technical outsourcing. The average Sinhala-speaking Sri Lankan will be able to understand English if they can translate and interpret Singlish to English or Sinhala and vice versa. If regular people employ excellent translation software, Sri Lanka, as a country and one of the world's top tourist destinations, may benefit from significant economic growth due to the high English literacy growth.

This paper introduces a translator which urban language translator called “FluentSL”. The main functionality of that application is to translate Sinhala to English and vice versa and translation of Singlish to English and Sinhala. Using 3rd party APIs which are already developed [4], and logical technical methodology of stemming would be used when developing the application. It is web web-based mobile friendly application that the user can use to input data using scanning a document and typing inputs. The application provides a dictionary so that the user can also be able to learn and get to know the words and their meaning better. The User gets suggestions when the words are typed [5].

The Main Objective of the research paper is to implement this application and get the public to use it as described earlier as a nation, Sri Lanka is one of the best minds in South Asia to be open to the world market so that can highly benefit economically since Sri Lanka has a good pool of skilled worker.

2. Literature Review

The research paper [6] and the current research paper consist of many different technologies. In the mentioned paper, the translation from English to Sinhala is carried out systematically, assuring linguistic accuracy and semantic fidelity. It begins with the English Morphological Analyzer, which carefully examines each word and extracts important morphological information such as parts of speech, tense, gender, number, and case. This research serves as the foundation for later translation. The Morphological Analyzer's tokenized words are sent on to the English Parser. It functions as a syntactic detective, decoding the structure of English sentences and recognizing subjects, verbs, and other elements. The Translator module is critical in facilitating a smooth linguistic shift by converting English base words into Sinhala counterparts using a bilingual dictionary. Following that, the Sinhala Morphological Analyzer generates appropriate Sinhala words based on previously collected grammatical information. It uses three dictionaries and inflection rules to create accurate Sinhala words. The Sinhala Parser then constructs a grammatically correct Sinhala sentence, considering syntactic norms and semantic nuances. The outcome is the Final Sinhala statement, which grammatically and semantically captures the core of the original English statement. Intermediate editing, which is optional, can improve translation quality by addressing ambiguities and complications. The voyage concludes with the Output, which is a translated Sinhala statement that bridges the linguistic divide between English and Sinhala, promoting cross-cultural understanding. This methodical technique, which employs specialized modules and occasional human interaction, lies at the heart of the English-to-Sinhala machine translation process, allowing for effective communication and connection between disparate language communities.

In the current research paper, the system's input processing phase is highly versatile, accepting text through various channels, including keyboard inputs, voice commands, and scanned documents. This adaptable system can handle individual words, complete sentences, or even entire paragraphs. However, adherence to specific language guidelines, including proper grammar and sentence-ending punctuation, is a prerequisite for successful translation. Upon receiving input, the system dives into sentence analysis, identifying sentences based on punctuation marks and meticulously tokenizing them into individual words. Notably, it distinguishes between English and Singlish words, a crucial step for accurate translation. In the database interaction stage, English words find their place in a data table, while Singlish words undergo stemming to discover matches within the MongoDB database. MongoDB's full-text search capabilities further enhance word retrieval, and vital word-related information such as part of speech and gender is meticulously recorded. The subsequent word combination process seamlessly assembles words into coherent sentences. Initially, the system doesn't consider specific patterns, but in a

subsequent step, it applies English grammar rules. Additionally, Singlish word postfixes play a pivotal role in determining sentence context. To ensure impeccable grammar, an external API is integrated for grammar checks and valuable suggestions, guaranteeing that the translations displayed maintain the highest level of linguistic accuracy. This comprehensive process efficiently bridges the gap between languages, delivering precise and contextually relevant translations [7].

The main difference between them is that the current research paper uses a defined stack to implement the system and the previously conducted research specifically mentions that it uses the dictionary to translate the words while using a system called "Morphological Analyzer" and "translation Engines". Thus, the current paper flows a simple process rather than a complex process.

3. Methodology

The system employs an Object-Oriented methodology for its design and leverages Visual Studio Code as its Integrated Development Environment (IDE). It relies on NodeJS as the programming language and MongoDB as the database, embodying the MERN Stack (MongoDB, ExpressJS, React, NodeJS) development technology. This choice is motivated by its speed, scalability, flexibility, user-friendliness, cost-effectiveness, and enhanced security. The system uses stemming as part of its algorithm, a crucial element in Natural Language Processing (NLP), aimed at reducing inflected word forms to their base form [8].

NPL also known as Natural Language Processing includes language libraries that are trained and use machine learning algorithms and artificial intelligence in order to translate any given text. In order to prepare words and text there are specific is techniques used. This is the technique of text normalization which consists of two main parts which are Lemmatization and Stemming [9].

Stemming is the process of splitting the end of each word in order to bring the word to the base form but the problem that tends to happen is that the word occasionally strays from the original meaning.

The representation (as indicated by in Fig. 1) shows the problem when it comes to using stemming. The alternative to fix this issue is Lemmatization which avoids this issue and does everything that is possible in stemming as well (as indicated by in Fig. 2).

Fig. 1 The stemming process

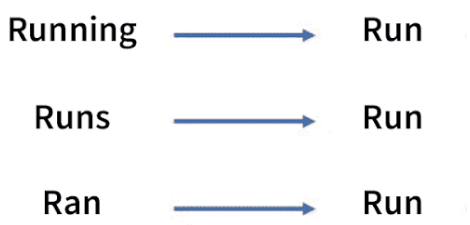

Fig. 2 The Lemmatization process.

In the initial phase of the translation process, the translation engine accepts input text through various means such as keyboard, voice, and scanned inputs. These inputs may consist of words, sentences, or paragraphs. To initiate the translation process successfully, the engine necessitates clear adherence to specific guidelines. Each sentence entering the system must be grammatically correct and end with a period (.), single quotation marks ('), double

quotation marks ("), or a question mark (?). Furthermore, words should be separated by spaces (), commas (,), or backslashes (/). Adhering to these guidelines is essential to ensure accurate sentence translation.

The translation engine then proceeds to analyze the input text. Upon detecting a period or question mark, it treats the text up to these marks as a single sentence. Each word from that sentence is added to a sentence object, which is then included in an ArrayList. An ArrayList is declared for each sentence. To illustrate this process, consider the sentence "apita heta school eke exams thiyenawa" being converted from Singlish to English, as depicted in Fig. 3.

Once the translation engine identifies a word as either English or Singlish, English words are added to the data table (denoted as 'E' in Fig. 3). For Singlish words, stemming is applied until a match is found in the MongoDB database. This search process utilizes MongoDB's full-text search mechanism [10], and upon locating the word, it is added to the data table. Each sentence corresponds to a table containing word-related information, including whether it is a noun or verb, gender, and other relevant details.

After gathering word information, the translation engine proceeds to combine words (as indicated by 'F' in Fig. 3). In this joining process, the initial pattern is not considered. However, in the second step, the engine applies English grammar rules [11], such as placing adjectives before nouns and using indefinite articles like 'a' or 'an.' Additionally, Singlish word postfixes play a role in sentence context determination ('G' in Fig. 3). Considering these rules and facts, the engine formats the sentence accordingly. For example, in the case of the Singlish postfix "eke" in the previous sentence, the system adds "at" before the school when the Sinhala word ends with "eke." Various English words are utilized for translating these Sinhala word postfixes, as outlined in TABLE *1*. Finally, the system employs an external API to perform grammar checks in the last step [12]. Also, the system will provide suggestions based on words which will enhance the user experience.

TABLE 1 SINGLISH SUFFIXES TO ENGLISH

Sinhalese	**Singlish**	**English**
ක්	k	a / an
කින්	kin	by a / by an
නක්	nak	on a / on an
උව	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
ආව	aava	the
එන්ව	nva	the
න්ට	nta	the
උට	uta	a / an
එගක්	eka	at

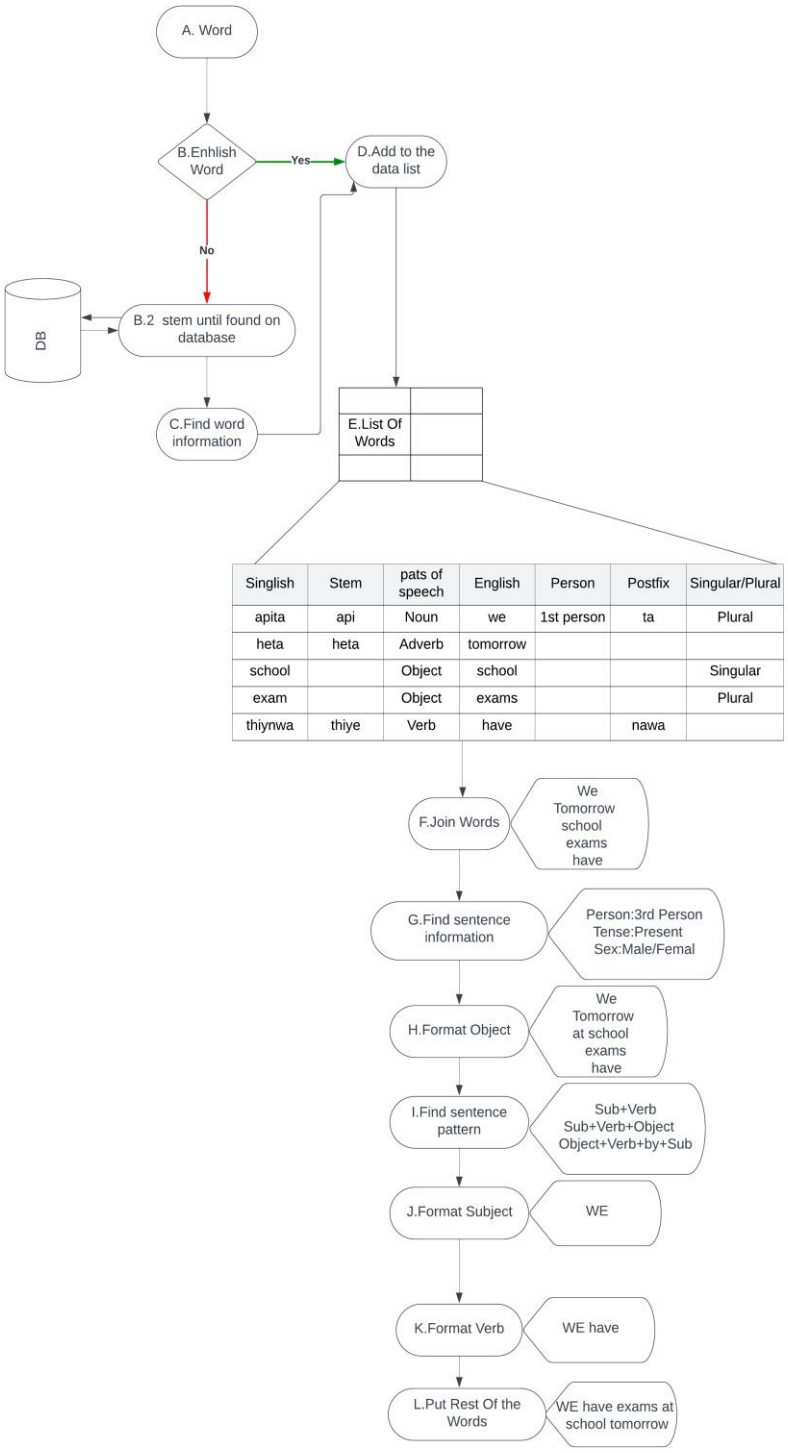

Fig. 3 Diagram of translation process

4. Existing systems

When it comes to existing systems many systems have been implemented throughout the years. Such as Google translator, hela kuru, and many more. There is quite a bit of translation apps out there due to the improvements in technology.

Google Translate is an application that is very popular throughout the world. Google Translator was first released in 2006 and is still one of the most useful tools of Google, which is used by many travellers and learners. Google Translate also comes with many features as well. The ability to translate whole documents or sentences, the ability to translate text through images, translate any spoken content, and also a very important feature of being able to translate whole applications which can be done through using Google's translation engine. And the rate of speed it translates only adds to the vast number of features that are included within the system. It does come with its limitations as well since the translation even though it is very good tends to have some occasional grammatical issues. And it also tends to be situations where some languages get better translations than other languages. Even though Google often time has the ability to identify and keep within the correct context there are occasions where it tends to stray out of the actual context as well [13].

Hela kuru even though is not prominent when considering a global standpoint, has become very prominent when considering there are about 10 million downloads who are Sri Lankans it's safe to say that their claims of being one of the best translation apps are no acceptable. When it comes to features this application has a handful of features. Starting from being able to get news and having multiple payment methods for bill payment reloads and an inbuilt dictionary as well. But all these features considered the most prominent feature is their custom keyboard. It is easy to use and efficient which is one of the reasons it became very famous in Sri Lanka. The only thing that's holding back this application is the fact that there tends to be buggy as reported in the review section of the users [14].

When considering applications related to translation there are many to be considered however out of all of those applications one noticeable one would be the translation application developed by the students at the University of Colombo School of Computing [15].

To elaborate the students of UCSC developed a translation engine that takes Sinhala written in English letters and turns them into pure Sinhala words with Sinhala letters.

As seen in the Fig. 4 The phrase "mama bath kawa" has been converted into "මම බත් කව". The difference between the proposed system and the UCSC translation engine is that the proposed system translates the sentences to English regardless of the sentence having Sinhala or English words rather than translating only Sinhala words. Regardless it is a translation engine that should be considered as a stepping stone for today's application related to translation.

Considering that there are so many translation applications the factor that makes the proposed system so unique is due to the ability to translate Singlish to English. The ability to translate Singlish to English comes in very handy since it is being used quite a lot in the present day and having the ability to translate will be very beneficial for education purposes and will be very convenient as well.

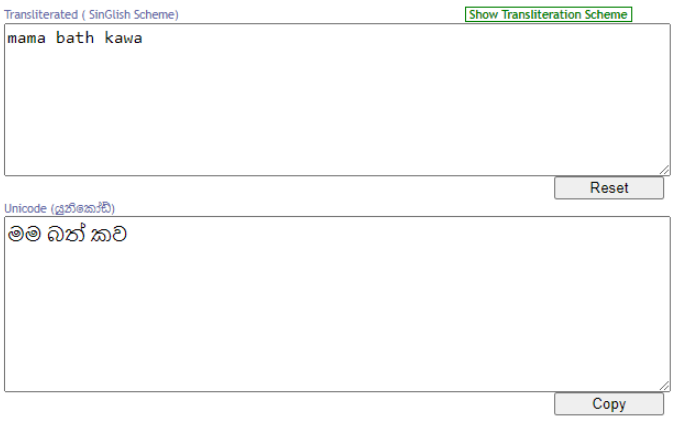

Fig. 4 Unicode Real Time Font Conversion Utility

5. The system and features

Apart from the main feature of translating Singlish to English which as mentioned is accomplished.

Through the process of stemming, which explains the process of getting each word to its base form. Other than the main functionality that's mentioned, there are more additional features that are being covered.

The best way to learn any language is to understand it and what better way to understand words that are hard to comprehend than using a dictionary? The main purpose of including a library is to increase the ease of learning. To achieve the functionality of an integrated library a React Dictionary is used. [16] This is a JavaScript library that uses an API to retrieve information such as meaning and examples for inputted words and it also provides a variety of other languages to translate as well but in this case English will be the only language that will be useful in this scenario.

When it comes to translation, there are many ways to go by it; one of those ways is to directly translate through a PDF document which is one of the features that's included when considering the features. To achieve this PDF-scrapper which is a node package is used [17]. PDF-scrapper simplifies the extracting text from a document and it also provides additional features such as the amount of pages that require to be extracted and error handling as well. After the text is extracted from the pdf document the text is taken through the systems main procedure to translate according to the user's need.

Another crucial part is the ability to suggest words while typing as seen in many other applications and keyboards, this feature has been a must-have feature within most apps that real related to language and typing. The fact that makes this important is people tend to make mistakes and this also assists the actual translation process when considering how the translation process is done as mentioned.

To achieve the best possible translation at the end of the translation before it's displayed the whole sentence goes through a grammar-checking process. This makes sure that any content that's being delivered is maintaining an exceptional level of grammar accuracy. To achieve this an API named "Grammar checker backed by OpenAI's GPT-3" [12] is used which analyses the given sentences to provide the errors. It also provides support for a variety of other languages as well.

Furthermore, a history of all the sentences that were translated is also maintained in the system. This is maintained through a Mongo DB collection which will store all the records in that collection and will retrieve them when needed. The user can delete any record as well. The history though simple serves a valuable purpose since it is quite important to be able to take a step back and go back to what has been translated beforehand.

These functionalities are included so that the user will have all the necessary functionalities to ensure that the user has everything necessary to learn and explore both languages which reflects the cultural diversity of both worlds.

6. Discussion

The purpose of this system is to enhance how people learn new languages. Singlish is something unique it's something that originates from two completely different languages however it is also a great way to learn since Sinhala is something that Sri Lankans are quite fluent in and using that in combination with English makes it easier to learn English since that gives a helping hand to the users. This system helps bridge the linguistic gap by increasing the ease of learning and exploring the English language.

While maintaining a 100 percent accuracy is impossible to get the most grammatically accurate translation there are measurements taken with grammar checkers, but it still is impossible to maintain that level of accuracy with the technology that is being currently used throughout this instance. However, with everything that's being used, there will certainly be an exceptional level of accuracy with the sentences that are translated.

Even though the current technology that is being used has its limitations as time goes on with new technological improvements the application accuracy can be taken to new heights especially when considering the improvements of artificial intelligence in the present-day proposed system could be taken to a level where it provides translations with higher accuracy.

7. Conclusion

In conclusion, this research represents the complexity of both languages as the linguistic gap between them is reduced it is important to recognize that the form of translation is ever-evolving and the methodologies that have

been used throughout this research will be improved as time passes by. The findings merely pave the path to an ever-evolving field.

This system serves as a way of assistance for non-English speakers and a steppingstone for individuals who wish to improve on this using the ever-evolving technologies to take the process of translation to newer heights.

References

[1] L. D.O, "Facts and Details," 2009. [Online]. Available: https://factsanddetails.com/south-asia/Srilanka/People_Srilanka/entry-7966.html.

[2] L. C. Zamora, "Terminology Coordination," 10 May 2020. [Online]. Available: https://termcoord.eu/2020/05/the-origins-of-the-english-language/#:~:text=English%20is%20a%20West%20Germanic,an%20English%20form%20of%20this.%20by%20Lidia%20Capitan%20Zamora(Journalist,%20web%20editor%20and%20social%20media%20expert).

[3] L. Wijerathna, W.L.S.L. Somaweera, S.L. Kaduruwana, Y.V. Wijesinghe, D. I. De Silva, K. Pulasinghe, S. Thelijjagoda, "A Translator from Sinhala to English and English to Sinhala (SEES)," Proc. International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, December 2012, pp. 14-18.

[4] o. sibari, "Rapid api," April 2023. [Online]. Available: https://rapidapi.com/sibaridev/api/rapid-translate-multi-traduction/.

[5] D. I. De Silva, P. K. D. A. Alahakoon, P. V. I. Udayangani, D. Kolonnage, M. H. P. Perera, and S. Thelijjagoda. "Application of Transfer based Machine Translations from Sinhala to English," In Proceedings of the 4th SLIIT Research Symposium, vol. 2, pp. 33- 36. 2008.

[6] B. Hettige and A. Karunananda, "Theoretical based approach to English to Sinhala machine translation," in *Industrial and Information Systems (ICIIS)*, Colombo, 2010.

[7] D. De Silva, A. Alahakoon, I. Udayangani, V. Kumara, D. Kolonnage, H. Perera, S. Thelijjagoda, “Sinhala to English Language Translator," 4th International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, 2008, pp. 419- 424.

[8] V. Welgama, "Evaluation of a shallow stemming algorithm for sinhala," colombo, 2011.

[9] Saumya, "Analytics Vidhya," Analytics Vidhya, 28 June 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/#:~:text=Stemming%20is%20a%20process%20that,form%2C%20which%20is%20called%20Lemma..

[10] "MongoDB," MongoDB, Inc, 2023. [Online]. Available: https://www.mongodb.com/basics/full-text-search.

[11] P. Herring, "The Farlex Grammar Book:," FARLEX International, 2016.

[12] N. Brodin, "github," GitHub, Inc., Augest 2023. [Online]. Available: https://github.com/NathanBrodin/grammar-checker.

[13] D. Adewusi, "Scientific Editing Blog," 5 March 2021. [Online]. Available: https://www.scientific-editing.info/blog/everything-you-need-to-know-about-google-translate/.

[14] B. L. (. Ltd., "Helakuru Superapp - Sri Lanka," 2023. [Online]. Available: https://apps.apple.com/us/app/helakuru-superapp-sri-lanka/id1012390370.

[15] T. R. L. -. U. o. C. S. o. Computing, "Unicode real time font conversion utility," University of Colombo School of Computing, 2006. [Online]. Available: https://ucsc.cmb.ac.lk/ltrl/services/feconverter/t1.html.

[16] P. Agarwal, "github," 4 July 2021. [Online]. Available: https://github.com/piyush-eon/react-dictionary-wordhunt/tree/master/.

[17] m. kozan, "gitlab," 2019. [Online]. Available: https://gitlab.com/autokent/pdf-parse.