Swa-Bhasha Dataset: Romanized Sinhala to Sinhala Adhoc Transliteration Corpus

T.G.D.K. Sumanathilaka
Department of Computer Science
Swansea University
Swansea, Wales, United Kingdom
deshankoshala@gmail.com

Nicholas Micallef
Department of Computer Science
Swansea University
Swansea, Wales, United Kingdom
nicholas.micallef@swansea.ac.uk

Ruvan Weerasinghe
School of Computing,
University of Colombo
Colombo 007, Sri Lanka
arw@ucsc.cmb.ac.lk

***Abstract*—In the context of a changing society and rapid technological advancements, the prevalence of social media platforms and instant messaging services has significantly strengthened the usage of native languages. In Sri Lanka, Sinhala and Romanized Sinhala have emerged as popular typing languages, owing to the widespread use of informal shorthand-based typing and internet acronyms for quicker communication. However, due to the limited availability of resources, linguistic support for these languages is limited, making them low-resource languages. To address this resource deficit, this study proposes the development of a rule-based transliteration tool that can annotate Sinhala words into Romanized Sinhala, accommodating the diverse ad hoc typing patterns used by the community. The research approach involved a comprehensive survey employing a stratified sampling method, considering variables such as age, gender, and language proficiency. 215 participants were presented with an online survey comprising 12 Sinhala sentences to capture various transliteration patterns related to Sinhala characters which are necessary for the annotation process. Analysis of the survey responses led to the formulation of 92 general rules and 26 special rules, encapsulating ad-hoc Romanized Sinhala typing patterns. Using these rules, Sinhala dictionaries were annotated, building a large corpus of data which consists of Sinhala and its Romanized Sinhala patterns. The annotated dataset was validated using a back transliteration tool, achieving an 84%-word accuracy rate. This innovative transliteration annotator can be used to mitigate the resource constraints associated with Sinhala to Romanized Sinhala transliteration. GitHub link:**
https://github.com/Sumanathilaka/Swa-Bhasha-Sinhala-Singlish-Dataset

Keywords—Annotation, Dataset Creation, Romanized Sinhala, Transliteration, survey

I. Introduction

With the advent of social technology, social media usage became prominent [1]. The young community use social media apps to share their day-to-day activities, ideas and suggestions on current situations and new trends. The quick accessibility to these networking platforms has shaped the communication trend to a new level. However native language speakers tend to use their languages in social media with the introduction of multi-language keyboards. This has been a great opportunity to communicate and share ideas with others irrespective of the language barrier. Sinhala is one of the most spoken languages (74%) in Sri Lanka and approximately 13 million people speak the language all around the world [2]. Sinhala is inherited from Sanskrit, and it is considered to be an Indo-Aryan Language. There is a significant amount of research conducted in Sinhala despite the lack of resource availability. Mainly Sinhala to English Translation [3,4], Sentiment analysis [5,6], Information retrieval [7], Summarization, Text to Speech Synthesis [8,9] have been implemented and tested. But most of these researches were focused on the Pure Sinhala Language. Code mixed languages, Singlish and language mutations were not researched much. The unofficial Language Singlish where Sinhala is written in Latin alphabets and a mix of English words, Romanized Sinhala representation of Sinhala is yet to be researched for the above-mentioned domains.

Transliteration and reverse transliteration especially have not been well implemented due to the lack of resource availability. Transliteration is the process of writing words or letters of a language using other familiar languages [10]. As an example, *ආයුබෝවන්* can be represented using latin characters "*Ayubowan*". But with the influence of shorthand typing or short acronyms with Texting based typing the word can be presented in many forms like "Ayubown, Aybowan, Aubowan". Different formats of writing have not been properly handled in many of the research when working with Romanized Sinhala text due to the unavailability of such datasets with different typing patterns. So as a solution to the scarcity of datasets which contain Sinhala and its Romanized Sinhala representation, the authors of this work have introduced the Swa-Bhasha Dataset which uses an automated way to annotate data based on the survey data. This approach mainly consists of five steps namely conducting a survey, segmentation and alignment, pattern identification, rule generation and automating the annotation. These steps are further elaborated in the below sections. As a result, of the above steps, 4 different forms of datasets have been released which can be used for Transliteration purposes, Word sense Disambiguation in Transliteration and word suggestions to handle Transliteration ambiguity.

In summary, this introduction chapter has provided a comprehensive overview of the research area, highlighting the key factors of this study. Moving forward, the subsequent sections will focus on related works, detail the methodology employed, present the experimented results, draw conclusions, and outline potential research works for future exploration. These sections aim to offer a deeper understanding of the subject matter and contribute significantly to the discourse on this topic.

The following key contributions can be highlighted.

1. Development of a Sinhala-Romanized Sinhala word-level dataset, Romanized Sinhala-Sinhala Social Media sentence pair dataset, and Romanized Sinhala Transliteration Dataset for Word Sense Disambiguation (WSD).

2. Introduction of a data annotation algorithm tailored for low-resource languages, leveraging survey analysis, segmentation, and alignment methodologies within the annotation process.

II. Related Works

A. Datasets

1) Dakshina Dataset by Google Research [11]

The Dakshina dataset is a comprehensive resource catering to 12 South Asian languages. This collection comprises textual content in both the indigenous script and the Latin alphabet. The dataset is structured to include native-script Wikipedia texts, a corresponding Romanization lexicon, and extensive sentence parallels in both the native script and the fundamental Latin script. There are 10000 Romanized Sinhala Sentences with approximately 14.3 native words per sentence. This has been based on Wikipedia. Also, 25000 Romanized Lexicons can be found in the dataset. This dataset has been used in many research activities related to Sinhala and Romanized Sinhala *[12]*.

2) Liwera Dataset on Social Media hate speech [13]

Liwera and his coauthors have created a social media-based dataset which mainly scraped from YouTube comments. This dataset has been used to train the Ngram model which transliterates Singlish to Sinhala. This dataset consists of 5000 sentences taken from political and news videos mainly focused on Hate speech. It has been arranged in the below format to train the Trigram model used in their research work.

palayan/පලයං boru/බොරු karaya/කාරය yanna/යන්න tho/තෝ

3) Xtreme Up [14]

The research team developed datasets for 88 less-commonly studied languages, aiming to improve various language technologies like speech recognition, text reading, and translation. One of these datasets focuses on Sinhala and is useful for reading text and converting it into Romanized Sinhala. It contains 1000 records for training and 120 records for testing. All the data is neatly organized in a JSON file, making it easy to find both the Sinhala text and its Romanized version. This setup simplifies how researchers and developers can access and use the information.

4) Singlish Data set of Sandaruwan [15]

This work utilized the IWSLT'15 English-Vietnamese parallel dataset, sourced from Stanford University, specifically extracting English sentences. These sentences were employed to produce Sinhala translations using the Google Translator API and establishing a Singlish dataset in parallel with the original English sentences through the Google Pronunciation API. Following data preparation, an additional script was devised to cleanse the information derived from the pronunciation API.This refinement of this dataset aimed to capture the essence of Singlish language patterns as closely as possible, aligning with natural writing conventions. This dataset consists of 0.26 M language pairs (Singlish- English).

Not only the above datasets but there are also some other datasets used in different transliteration systems. Lahiru *et al.* [16] have come up with 10000 Singlish tweets and use an error correction module in the transliteration model to avoid the different transliteration types. However, the system is not capable of handling the ambiguities in Singlish to Sinhala Transliteration due to this limitation. Rushan *et al.* use the LTRL dataset from the Language Technology Research Laboratory which consumes WhatsApp messages and the Dhakshina dataset which is mentioned above. Athukorala *et al.* trained her model using a word corpus with approximately. 200k Sinhala words which have been extracted from a 10M work corpus licenced under the University of Colombo School of Computing.

TABLE I. COMPARATIVE ANALYSIS OF DATASETS

Dataset Name	**Size**	**Nature**
Dakshina Dataset [11]	0.01 M Sentences and 0.025 M Romanized Lexicons	Sentence Pair Word pairs
Liwera Dataset [13]	0.005 M Sentences	Sentence Pairs
Xtreme Up [14]	0.001 M Sentences	Sentence Pairs
Sandaruwan *et al.* [15]	0.26 M Singlish English Sentences	Sentence Pairs
Swa-bhasha Dataset	0.004 M Sentences 7.13 M Romanized Lexicons	Sentence pairs Word pairs

B. Transliteration systems using the above datasets.

Presently, several commercial Transliterators such as Helakuru [17] and UCSC Transliterator [18] employ a rule-based methodology. Silva and Ahangama's recent study integrate a module that corrects errors with a rule-based method for converting Singlish to Sinhala [16]. Khan's Hinglish to Hindi transliterator also utilizes a similar approach but cannot handle ad hoc transliterations [19]. Liwera et al introduced a Trigram based module in conjunction with the rule-based approach, achieving a word level accuracy of 0.64 [13]. A system that uses numerical mapping and a fuzzy logic-based method with a unigram tagger achieves a 0.74 word-level accuracy. This was Swa-bhasha initial work which was introduced by Athukorala et al. [20]. Sandaruwan's recent study employs the Seq2seq model with Bi-directional LSTM and an attention layer for the encoder and Unidirectional LSTM for the decoder [15]. Liwera et al.'s study implements a hybrid approach for transliteration [13]. Sumanathilaka et al. incorporate an Ngram model, rule base, and Trie structure for word suggestion, enhancing machine transliteration effectiveness with 0.84-word level accuracy and 0.93 character level accuracy [1]. Commercial tools like the Google input tool [20] integrate a Knowledge and Neural model to perform the transliteration.

III. Methodology

According to the recent works and datasets examined in the previous section, available datasets cannot be used to

handle ad-hoc transliteration as many of these datasets are mainly focused on a limited level of writing patterns.

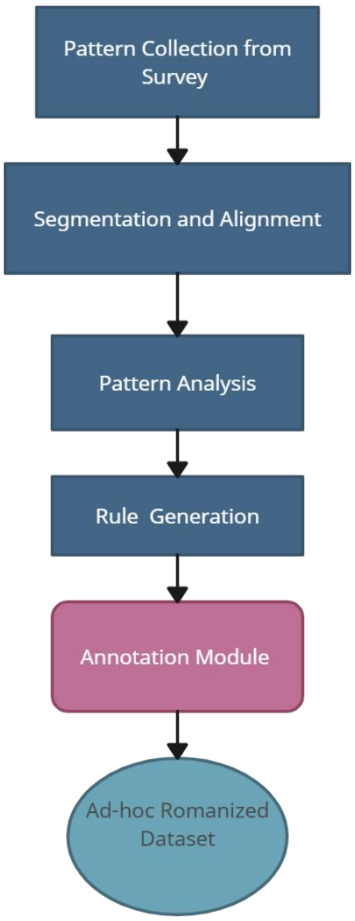

Fig. 1. Dataset Creation Activity Flow

So as a solution to this limitation, the Swa-Bhasha dataset has been introduced using a rule and survey-based approach. The dataset creation step can be categorized into the steps below.

A. Data Collection Approach

A dataset comprising "Romanized Sinhala sentences" and their corresponding Sinhala sentences sourced from social media content has been identified. Additionally, the Dakshina dataset obtained from Google [11], containing transliterations, has be employed to enrich the transliteration set necessary to enrich the corpus. The author of this study has collected a dataset which includes diverse categories sourced from YouTube's social media content. This dataset contains a broad spectrum of Romanized Sinhala and respective Sinhala words derived from various

Social Media platforms. The data extraction from YouTube has been facilitated using the freely available Face Pager [21] application, specifically focusing on sentences containing Romanized Sinhala. The dataset creation encompasses themes such as social backgrounds, music, politics, news, religion, and technical topics. Subsequently, the collected dataset was manually transliterated to generate specific Sinhala sentences related to the content.

B. Survey Analysis

The researchers distributed online surveys [22] across various communities to gather information on typing patterns. The initial survey was conducted for nearly 1.5 months. By analyzing inputs from 215 users, authors identified diverse typing styles, which were then studied to establish rules for the Data annotation algorithm. Subsequently, a subset of selected sentences was reintroduced to the selected 25 participants to observe any alterations in their typing habits over time. The second round was conducted 6 months after the initial survey. The researchers employed stratified sampling to target specific audience groups and study their corresponding typing patterns. The participants were between the age of 18-40 and frequently used English keyboards to type Romanized Sinhala. A few instances from survey questionnaire can be found in TABLE II.

TABLE II. SINHALA SENTENCES USED IN THE SURVEY

විවිධ අවස්ථාවල දී ක්ෂුද්‍ර ජීවීන් පාලනය සහ ජීවානුහරණය කිරීම සඳහා නොයෙකුත් ශිල්ප ක්‍රම භාවිතා කරනු ලැබේ.
මිලේච්ඡ ලෙස සාතනය කර තිබූ මිනිය සැඟවීමට උත්සාහ නොකර කෲර ලෙස විසිකර දමා තිබූ අයුරු දිස්විය
ඓතිහාසික වටිනාකමකින් යුක්ත වූ ඒ මහා බෝධීන් වහන්සේ කපා දැමීමට තැත් කළ සියලු දෙනා හට ස්වභාවධර්මයා විසින් දඬුවම් ලබා දුනි
ප්‍රථම වතාවට පන්සල්ට පිය නැගූ පුංචි ෆාතිමා සන්ටාරය නාද කොට ඉන්පසුව ඊතලයක් සේ දිව ගියේ බුදු මැදුරටය.
දේශපාලන හැලහැප්පීම් මැද පලිගැනීම් පමණක් අරමුණු කරගත් සමාජයක, කන්නන්ගර මහතා නිදහස් අධ්‍යාපනය හඳුන්වා දෙමින් කළ සේවාව ඉමහත්ය.
කවියෙකුගේ කවි සිතුවිල්ලකට ගීත රචකයෙකුගේ ලයාන්විත ගී පබැඳුමකට සිත්තරෙකුගේ චමත්කාර සිත්තමකට හේතුවූ බොහෝ පාරිසරික පසුබිම් දැකගත හැක
ඇත අතීතයේ සිට පැවත එන සංස්කෘතියේ එක් විශේෂාංගයක් ලෙස රජදරුවන් විසින් තැනූ වැව් සහ දාගැබ් හෙළ සිංහලයාගේ විශ්මිත හැකියාවන් පිළිබිඹු කරයි.
බටහිර වෛද්‍ය විද්‍යාවේ ඇතැම් හිඩැස් උෆන පූර්ණය කිරීම් සඳහා ආයුර්වේද වල යොදා ගන්නා ඖෂධ සෘජුවම ඕනෑ වීම ඒවායේ ඇති ඵලදායි බව පැහැදිලි කරයි.
කොම්පඤ්ඤ වීදියේ පුංචි සිඤ්ඤෝ, මඤ්ඤොක්කා අරන් කඩමණ්ඩියට යන ඥානපාල ට කතා කළේ විස්කිරිඤ්ඤා දෙකක් ගෙන්න ගන්න.

The survey was conducted in two time phases to identify the changes in typing patterns with time. During the initial round, an online survey containing 12 Sinhala sentences covering the Sinhala alphabet and its consonant and vowel patterns was shared with the participants. The 12 sentences have been arranged in a manner of covering all the characters in the Sinhala Alphabet. The users were given the freedom to use their familiar device to fill in the form and more than 90% used their mobile phone and inbuilt keyboard to complete the survey. The second phase of the survey was

collected after the 6 months of the initial round. 25 participants were selected randomly from the initial survey for this phase. The 3 sentences from the initial survey have been shared with the participants. This has been randomly selected to avoid the biasness in the study. The results depict that typing patterns change concerning the time and the mood of the user. The below input from the same user in different time frames proved it.

Sinhala sentence: විවිධ අවස්ථාවල දී ක්ෂුද්‍ර ජීවීන් පාලනය සහ ජීවානුහරණය කිරීම සඳහා නොයෙකුත් ශිල්ප ක්‍රම භාවිතා කරනු ලැබේ

Romanized Sinhala Sentence in phase 1 : Wiwida awasthawala di kshudra jiwin palanaya saha jiwanuharanaya kirima sandaha noyekuth shilapa krama bhawitha karanau labe.

Romanized Sinhala Sentence in phase II : Wiwida awasthawala di kshudra jeewin palanaya saha jeewanuharanaya kiriima sadaha noyekuth shilpa krama bhawitha karanu labe.

C. Rule Generation using Segmentation and Alignment

Based on the above inputs from the survey users, sentences were segmented and aligned to identify the key differences in the writing patterns. Character level segmentation was performed, and Rule-based General Transliteration was used as the base sentence for the alignment. Achieved knowledge in Table IV was identified using this technique.

K	O	H	O	M	A	D	H	A
K	O	H	O	M	-	D	-	A
K	O	H	O	M	A	D	-	A
K	-	H	-	M	-	D	-	A

Fig. 1. Alignment of word කොහොමද

D. DataSet Annotation

The patterns recognized in the data collection phase served as the base for developing a data annotation algorithm capable of generating Romanized Sinhala words with various typing patterns. Specific patterns, such as vowel arrangements, consonant-vowel usage, and mapping English consonants to different Sinhala characters, were identified and formulated into rules as shown in Table IV.

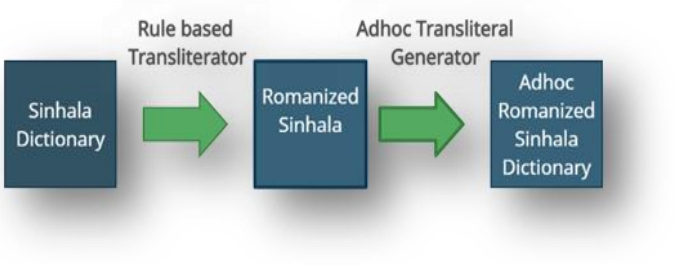

Fig. 2. Flow of Data Annotation Process

The rule-based transliterator, responsible for converting Sinhala to generalized Romanized Sinhala, operates on 60 rules for vowels and consonants, 18 rules for "hal" symbols, and 18 rules for special characters. The resulting generalized Romanized Sinhala words were inputted into the ad hoc transliteral generator, which consists of 12-character pattern rules, 6 rules for vowel combinations, and 8 rules for special characters. This generator created an ad-hoc Romanized Sinhala corpus which is known as Swa-Bhasha. To create a fully annotated dataset, the annotation algorithm, built on the established rules, was paired with a Sinhala Dictionary LTRL. The data annotation process flow is illustrated in Fig. 2, while the rules governing the ad-hoc Transliteral generator are outlined in Table V.

E. Preprocessing

The annotated dataset has been pre-processed to identify the redundant data. The duplicates have been removed and word ambiguity has been cross-checked manually to avoid inaccurate results. Using the above approach 7,134,803 (7.13 M) words were generated for unique 440,024 Sinhala Words. This corpus covers most of the possible typing sequences of the general Sinhala words. Table III presents the annotated sequences for a given Sinhala word.

TABLE I. RESULTS OF DATA ANNOTATION ALGORITHM

කොහොමද	khmda,khmad,khmda,khmad,kohmda,kohmad,kohmada,kohmad,khomd,khomda,khomad,khomada,khamda,khamd,khamad,khamada
කියන්න	kynn,kynna,kyann,kynn,kiynna,kiyann,kiyanna
පමණක්	pmnk,pmnak,pmank,pamnk,pamnak,pamank,pamanak
යන්න	ynn,yanna,ynna,yann

TABLE IV. USER INPUT ANALYSIS AND ACHIEVED KNOWLEDGE

Sinhala Sentences	Selected User Inputs from Users	Achieved knowledge
විවිධ අවස්ථාවල දී ක්ෂුද්‍ර ජීවීන් පාලනය සහ ජීවානුහරණය කිරීම සඳහා	Wiwida awasthawala di kshudra jeewin palanaya saha jeewanuharanaya kiriima sadaha noyekuth shilpa krama bhawitha karanu labe Wiwida awasthawaladi kshudra jeewin palanaya kireema sandaha noyekuth shilpa krama bhawitha karanu lebe.	Letter-wise transliterals identified. Ambiguity letters were identified.

නොයෙකුත් ශිල්ප ක්‍රම භාවිතා කරනු ලැබේ.	Wiwidha awasthawala Dee kshudra jeewiin paalanaya Saha jeewanuharanaya kireema sandaha noyekuth Shilpa krama bhawitha karanu labe. wiwida awasthawaladi kshdra jeewin paalnaya saha jeewanuharanaya kereema sadaha noyekuth shilpa krama bhawitha karanu lAbe. Vivida awastha waladi ikshudra jiwin palanaya saha jivanuharanaya kirima sadaha noyekuth shilpa krama bawitha karanu labe. wiwida awasthawaladi kshudra jeewin palanaya saha jeewanuharanaya kirima sadaha noyekuth shilpa krama bawitha krnu labe Vivida awastha waladi kshudra jeeween paalanaya saha jeewanuharanaya kireema sandaha noyekut shilpa karama baawitha karanu labe wiwida awasthawaladhii shrudhdha jiwiin paalanaya saha jiiwaanuharaNaya kiriima sadhahaa noyekuth shilpa krama bhaawithaa karanu lAbee. wiwida awasthawaladi kshudhra jeewin paalanaya saha jeewanuharanaya kireema sadaha noyekuth shilpa krama bhawitha karanu labe.	ව් can be represented using V or W. Vowel combinations identified.

TABLE II. RULES USED FOR DATA ANNOTATION ALGORITHM

Rules Set	Rules description
12-character patterns	• ඇ - A/E • ඈ - Aa/Ae • එ - e/a • ඊ - i/e/ee • කි(hal) – ki/ke • ඕ - oo/o • ච - cha/ca • ද - dha/da • ත - tha/ta • ෂ - sa/sha • ශ - sha/ sa • ව - wa/ va
6 vowel Combination	• a, e, i, o • ae, ea
8 special character rule	• බ - ka • ණ - na • ධ - da/ dha • ෆ - f • භ - ba / bha /Bha • ඹ - ba / bha • ළ - la • ඵ - ta

IV. Experimented Results

The dataset has been evaluated with Swa-bhasha Transliterator for Romanized Sinhala to Sinhala Transliteration [1]. The proposed Swa-bhasha transliterator is based on a hybrid model which uses the Ngram model followed by a rule-based model with Trie structure as a suggestion module. Both the N-gram module and suggestion module have been trained with the annotated data of Swa-bhasha corpus. The results generated using the model are significantly more accurate compared to the existing products/research in the literature. A precise summary of the dataset usage can be found in the paper by Sumanathilaka *et al.* [1]. The model was evaluated using the test data from Liwera and Dhakshina datasets which combinedly had 750 tuples. 84%-word level accuracy and 93%-character level accuracy have been achieved. The same methodology has been followed to create a Tamil Corpus with ad hoc Transliterations and it has been evaluated with a different transliterator known as Tamzi [24]. The above results confirm that the annotated data has positively impacted the improvement of the robustness of the results of the Transliterators.

V. Conclusion and Future Works

The Romanized Sinhala – Sinhala Transliteral based data limitation is a huge issue for effective literature research. The unavailability of proper Romanized Sinhala datasets has been a major challenge in many research. As Sinhala is known to be a low-resourced language, the availability of credible data resources is minimal. To overcome this issue, the authors of this work have introduced Swa-Bhasha corpus which contains a large amount of annotated data for Transliteration purposes. The dataset can be obtained from the below link for further research in the area. GitHub link: **https://github.com/Sumanathilaka/Swa-Bhasha-Sinhala-Singlish-Dataset**

As a future work, preprocessing can be automated to increase the efficiency of the annotation process. The built dataset can be tested with WSD for Romanized Sinhala Transliteration to further verify its credibility and usability. The data gathering chapter of the survey with the age above 45 years group was challenging due to the limitation of technical fluency. The utilization of longitudinal time series in data collection phases has introduced several challenges pertaining to the continuity of contact with the same resource individuals for the study.

The suggested annotation algorithm based on surveys can serve as a tool to interpret datasets from various low-resource languages, thereby extending the scope of this study. This solution would be a great initiative to overcome the language resource issue mainly with low resource languages.

Acknowledgement

The authors express their gratitude to the survey participants whose voluntary contributions were fundamental to this study. They acknowledge the invaluable insights provided by Sinhala linguistic experts, enriching the research. Special thanks are extended to Ms. H. Guruge for her expertise in formulating the survey questions, which significantly shaped this work. Their collective support and expertise were commendable in the completion of this study, and the authors deeply appreciate their contributions.

References

[1] T. G. D. K. Sumanathilaka, R. Weerasinghe, and Y. H. P. P. Priyadarshana, “Swa-Bhasha: Romanized Sinhala to Sinhala Reverse Transliteration using a Hybrid Approach,” in *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka: IEEE, Feb. 2023, pp. 136–141. doi: 10.1109/ICARC57651.2023.10145648.

[2] P. P. Singh *et al.*, “Reconstructing the population history of the Sinhalese, the major ethnic group in Śrī Laṅkā,” *iScience*, vol. 26, no. 10, p. 107797, Oct. 2023, doi: 10.1016/j.isci.2023.107797.

[3] L. Wijerathna *et al.*, “A Translator from Sinhala to English and English to Sinhala (SEES),” in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, Colombo, Western, Sri Lanka: IEEE, Dec. 2012, pp. 14–18. doi: 10.1109/ICTer.2012.6421408.

[4] D. De Silva *et al.*, “Sinhala to English Language Translator,” in *2008 4th International Conference on Information and Automation for Sustainability*, Colombo, Sri Lanka: IEEE, Dec. 2008, pp. 419–424. doi: 10.1109/ICIAFS.2008.4783983.

[5] S. R. Jamal and D. K. Sumanathilaka, “Extended Abstract: Emotional Analysis of News Using Affective Computing and Sentimental Analysis,” *Int. Conf. Innov. Info-Bus. Technol. Colombo Febr. 2022*, p. 3, 2022.

[6] B. Sharounthan, D. P. Nawinna, and R. De Silva, “Singlish Sentiment Analysis Based Rating For Public Transportation,” in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India: IEEE, Jan. 2021, pp. 1–7. doi: 10.1109/ICCCI50826.2021.9402548.

[7] S. Ranathunga, F. Farhath, U. Thayasivam, S. Jayasena, and G. Dias, “Si-Ta: Machine Translation of Sinhala and Tamil Official Documents,” in *2018 National Information Technology Conference (NITC)*, Colombo: IEEE, Oct. 2018, pp. 1–6. doi: 10.1109/NITC.2018.8550069.

[8] P. Jayawardhana, A. Aponso, N. Krishnarajah, and A. Rathnayake, “An Intelligent Approach of Text-To-Speech Synthesizers for English and Sinhala Languages,” in *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, Kahului, HI, USA: IEEE, Mar. 2019, pp. 229–234. doi: 10.1109/INFOCT.2019.8711051.

[9] N. Prasangini and H. Nagahamulla, “Sinhala Speech to Sinhala Unicode Text Conversion for Disaster Relief Facilitation in Sri Lanka,” in *2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, Colombo, Sri Lanka: IEEE, Dec. 2018, pp. 1–6. doi: 10.1109/ICIAFS.2018.8913360.

[10] “Transliteration,” *Oxford american Dictionary*. Oxford University Press, 2010.

[11] B. Roark *et al.*, “Processing South Asian Languages Written in the Latin Script: the Dakshina Dataset,” p. 11, 2020.

[12] R. Nanayakkara, T. Nadungodage, and R. Pushpananda, “Context Aware Back-Transliteration from English to Sinhala,” in *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka: IEEE, Nov. 2022, pp. 051–056. doi: 10.1109/ICTer58063.2022.10024072.

[13] W. M. P. Liwera and L. Ranathunga, “Combination of Trigram and Rule-based Model for Singlish to Sinhala Transliteration by Focusing Social Media Text,” in *2020 From Innovation to Impact (FITI)*, Colombo, Sri Lanka: IEEE, Dec. 2020, pp. 1–5. doi: 10.1109/FITI52050.2020.9424880.

[14] S. Ruder *et al.*, “XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages.” arXiv, May 24, 2023. Accessed: Nov. 27, 2023. [Online]. Available: http://arxiv.org/abs/2305.11938

[15] D. Sandaruwan, S. Sumathipala, and S. Fernando, “Neural Machine Translation Approach for Singlish to English Translation,” *Int. J. Adv. ICT Emerg. Reg. ICTer*, vol. 14, no. 3, p. 36, Jul. 2021, doi: 10.4038/icter.v14i3.7230.

[16] L. de Silva and S. Ahangama, “Singlish to Sinhala Transliteration using Rule-based Approach,” in *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, Kandy, Sri Lanka: IEEE, Sep. 2021, pp. 162–167. doi: 10.1109/ICIIS53135.2021.9660744.

[17] D. P. Bhasha, “Helakuru Transliterator.” Jan. 01, 2011. [Online]. Available: https://www.helakuru.lk/

[18] Language Technology Research Laboratory - University of Colombo School of Computing, “Uniocode Real Time Font Conversion Utility.” University of Colombo School of Computing, 2006. [Online]. Available: https://ucsc.cmb.ac.lk/ltrl/services/feconverter/t1.html

[19] R. Khan, “Hinglish 2 Hindi Converter,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 3, pp. 2530–2535, Mar. 2018, doi: 10.22214/ijraset.2018.3410.

[20] Google, “Google Translator.” Apr. 28, 2006. [Online]. Available: https://translate.google.com/

[21] T. K. Jakob Jünger, “Facepager 3.6.” 2014. [Online]. Available: https://facepager.software.informer.com/

[22] Deshan sumanathilaka, “A Survey to Collect Romanized Sinhala Typing Patterns.” [Online]. Available: https://docs.google.com/forms/d/1f0bkuMwU56JqJrdF2OhQiZbcNmh6B_3byeMxPJu_3ys/edit

[24] Anuja Herath, T.G.D.K. Sumanathilaka, “TAMழி: shorthand romanized tamil to tamil reverse transliteration using novel hybrid approach,” in *2023 23rd International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka: IEEE, in press.