Universal Sinhala Library: Language Specific Encryption Platform for Sinhala Language

Conference Paper · February 2022

DOI: 10.1109/ICMRE54455.2022.9734085

2 authors, including:

Dinindu Koliya Harshanath Webadu Wedanage
University of Wollongong
4 PUBLICATIONS **2** CITATIONS

2022 8th International Conference on Mechatronics and Robotics Engineering (ICMRE) | 978-1-6654-8377-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMRE54455.2022.9734085

Universal Sinhala Library: Language Specific Encryption Platform for Sinhala Language

Dinindu Koliya Harshanath Webadu Wedanage
Smart Infrastructure Facility
University of Wollongong
Wollongong, Australia
dkhww937@uowmail.edu.au

Samantha Thelijjagoda
SLIIT Business School
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
samantha.t@sliit.lk

***Abstract*—Security has become a significant challenge in the modern world. Number science and mathematics opened up a vast path to work with programming models to create innovative mechanisms which improve text encryption. Universal Sinhala Library is a Sinhala language-specific text encryption platform. Platform architecture has been designed to demonstrate every possible combination of the Sinhala alphabet, including comma, space and period. The hypothetical architecture includes every book that ever has been written in Sinhala, and every book that ever could be, including every poem, every scientific paper and every piece of document in Sinhala. The main goal of the research is to create an encryption mechanism for the Sinhala language. Linear congruential generator and extended euclidean algorithm have been used along with the Hull–Dobell Theorem to outline the backbone of the encryption platform. At present, it contains all possible combinations of Sinhala characters virtually. Sinhala text to be encrypted should be searched in the platform, and it will return the location of that particular text in the virtual library architecture, which is the encrypted text string for the searched Sinhala text. It is helpful to the people who are using or working with the Sinhala language, moreover for sharing and transferring Sinhala context securely.**

Keywords—text encryption, Sinhala language, security, language-specific encryption, virtual architecture

I. INTRODUCTION

Security is one of the biggest concerns of any text transferred through the internet. The growing capacity of digital encoding has opened up vast new avenues for archiving and distributing texts in virtual space, prompting many to declare the imminent obsolescence of print media, the book included. An exciting correlate to this situation is the revival of interest in and support for the idea of the universal Sinhala library, a virtual collection of every text in existence, albeit imagined as an immense database of digitized material with online accessibility. The core algorithm has to engage with every Sinhala character in the Sinhala alphabet. There are 60 contemporary Sinhala characters, and it contains 20 vowels letters and 40 consonant letters. The core algorithm demonstrates a virtual architecture named "Universal Sinhala Library", which includes each possible combination uniquely identified by its location in the virtual architecture. This particular location path identifies as the derived encrypted string.

There are various encryption algorithms widely used for security purposes. Image encryption, video encryption and text encryption can be categorised as subsections under encryption. Some algorithms are specified for a subsection, while some algorithms could be used exclusively. For example, the RC5 encryption algorithm is a fast symmetric block cipher fitting for hardware and software implementations. [1] Natural language processing-based text encryption mechanisms are a vital approach under text encryption [2]. A novel symmetric text encryption algorithm based on a logistic map has been introduced for high-security test encryption, and it can be used in real-time applications [3]. Text encryption has been employed in data compression, which has a significantly improved compression ratio [4]. Cryptographic approaches such as Elliptic Curve Cryptography have also been used to implement text encryption, allowing users to encrypt or decrypt any script with ASCII values [5]. Another ASCII value-based data encryption mechanism was introduced using symmetric key encryption technique [6]. Most of these encryption mechanisms has been derived from cryptography concepts and has not focused on introducing language-specific text encryption mechanism. The use of virtual architecture in an encryption mechanism has not been a concentrated concept in the domain.

Jorge Luis Borges has been introduced a virtual library in his book named "The Garden of Branching Paths", Virtual library includes every possible phrase in English that has been drafted and has needed to be drafted [7]. He introduced the virtual library as an endless number of hexagons; each hexagon has two facing doors, one door to enter the hexagon and another door to exit the hexagon, while the other four sides of the hexagon contain bookshelves, five to each side (Fig. 1).

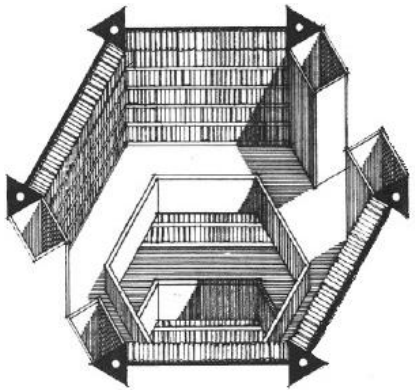

Fig. 1. Hexagon architecture.

Both doors of the hexagon lead to another hexagon. Each hexagon has a hexagon on top of it and a hexagon below it. A spiral staircase located at exiting and entering doors between two adjacent hexagons allows one to walk to the upper level of the library and the lower level of the library [7] (Fig. 2). The library of babel is a website designed by Jonathan Basile. It is a demonstration of Borges' library of babel explained in his book. Currently, it contains all possible pages of 3200 characters, about 10^{4677} books. He has also designed an image archive that contains all possible images that could ever exist [8].

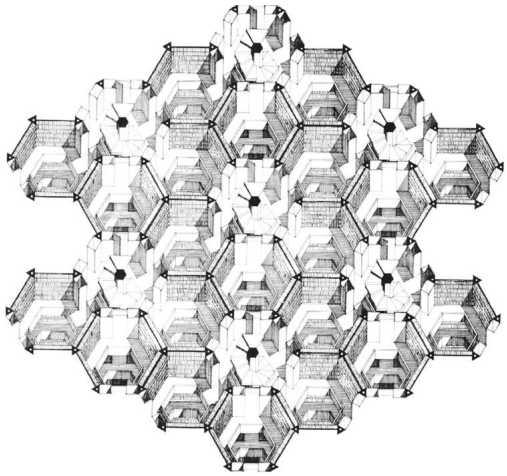

Fig. 2. Floor plan of Borges' Virtual Library, Library of Babel.

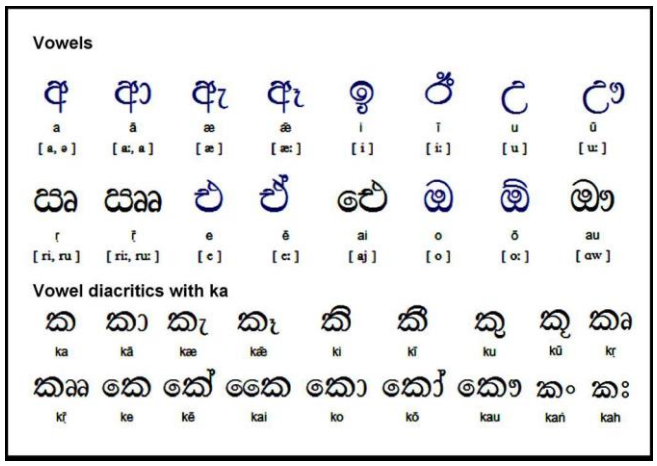

Fig. 3. Sinhala vowels and vowel diacritics.

The Sinhala language is derived from the Indo-Aryan group of languages [10]. The Sinhala language has 60 letters, including 20 vowels and 40 consonants (Fig. 4). Apart from that, the Sinhala language has 19 characters which assist in using the Sinhala language. Each consonant comes together with a vowel and forms a vowel diacritic [9] (Fig. 3) Sinhala to English translation demonstrates the Sinhala letters, Singlish letters which means writing Sinhala using English letters and English translation. This demonstration clearly illustrates the formation of vowel diacritics and their use [10].

Universal Sinhala Library is a successful attempt to design a Sinhala language-specific encryption platform inheriting the Borges' imaginary virtual library and Basile's library of babel website. Primary objectives include design an algorithm to work with Sinhala characters and encrypting Sinhala language content securely. The paper is structured as the introduction in the first section, showcasing the related work and research gaps. Section two demonstrates the overall methodology of the encryption mechanism, the third section illustrates the results, and the last section concludes all sections.

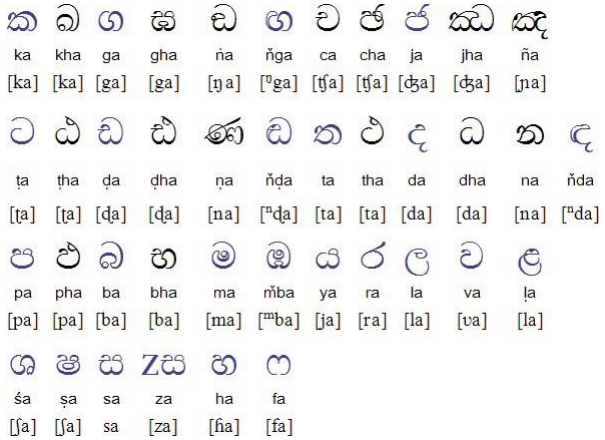

Fig. 4. Sinhala consonants

II. Methodology

We developed Universal Sinhala Library as a web application where users can search any Sinhala content or upload a softcopy of any Sinhala content for encryption. Initially, it was designed as mentioned in Borges' virtual architecture, and Basile's library of babel [7] [8]. Later on, we made it customizable where users can customize the initial architecture as they prefer. The implementation of the platform is divided into three major components as below.

- Developing base conversion algorithms fitting for the Sinhala language
- Developing the pseudo-random number generation algorithm
- Developing an algorithm to break down Sinhala contentinto a form where the core algorithm supports.

The above three algorithms work together in the web application; accordingly, all the above algorithms collectively form the core algorithm. The core algorithm functions as a decryption method; therefore, it has a mirror algorithm, which we named the inverted algorithm used as the encryptionmethod.

The Sinhala language has 82 characters, including vowels, consonants, assisting characters, space, comma and period. In order to represent Sinhala content as a number, each character should be given a number; the algorithm is designed to take one page at a time, which means 3200 Sinhala characters; then, any Sinhala content could be represented as a base-82 number. That number is converted into a base-10 number which is fed into the inverted algorithm. The inverted algorithm produces a base-10 unique

output number for the input seed number. That unique base-10 number is broken down into two parts defined in the virtual library architecture: hexagon number and page location. Page location consists of wall number, shelf number, volume number and the page number. Finally, the hexagon number is converted into a base-36 number. The concatenation of the hexagon number and the page location is represented as the encrypted text (Fig. 5).

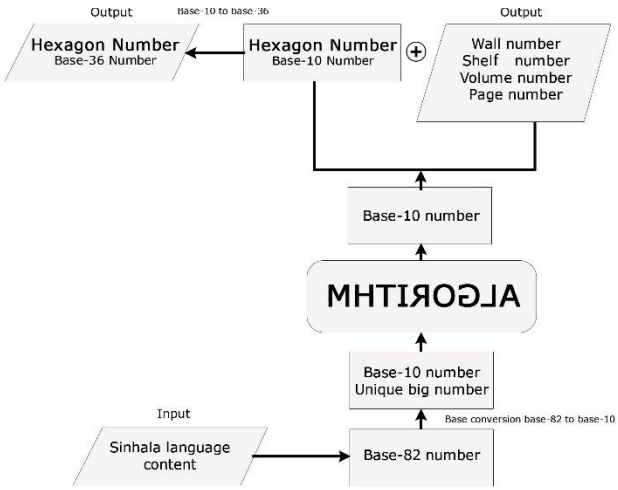

Fig. 5. Encryption algorithm flow chart.

The core algorithm, which is the mirror algorithm of the inverted algorithm, has been employed to decrypt the text. It takes base-36 hexagon number and base-10 page location as inputs. Base-36 hexagon number turned into base-10 number and consolidated base-10 number is taken as the input seed.

The core algorithm produces a unique base-10 number, which is converted into a base-82 number. Then base-82 number is turned into Sinhala content (Fig. 6).

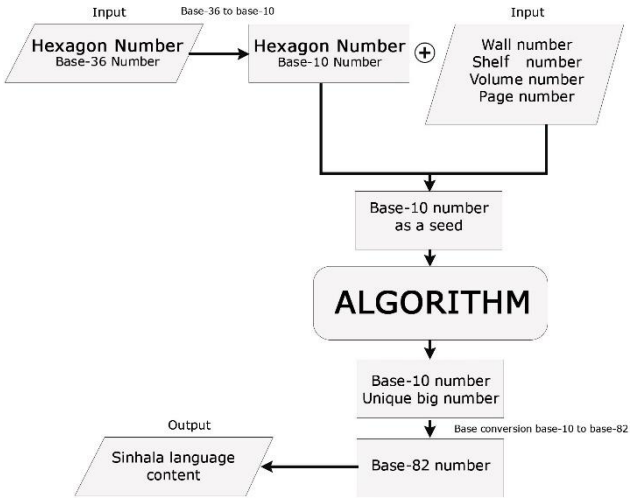

Fig. 6. Decryption algorithm flow chart

The linear congruential generator is used in the core algorithm to yield a sequence of pseudo-randomized numbers determined with a discontinuous piecewise linear equation as shown in equation (1). Recurrence relation has been used to design the pseudo-random number generator [11].

$$X_{n+1} = (aX_n + c) \mod m \quad (1)$$

Where, X is the sequence of pseudorandom values, the *modulus* (m) , $0 < m$, the *multiplier* (a), $0 < a < m$, the *increment* (c), $0 \quad c < m$, and the *seed* (X_0), $0 \quad X_0 < m$ are constant integers which defines the pseudorandom number generator.

When $c \neq 0$, accurately selected parameters allow a period equal to m, for all seed values. This will occur if and only if: (1) m and c are relatively prime (2) a-1 is divisible by all prime factors of m, and (3) a-1 is divisible by 4 if m is divisible by 4. These three requirements are referred to as the Hull–Dobell Theorem.

As shown in equation (2), the extended euclidean algorithm is used to find the accurate coefficients for the above algorithm.

$$ax + by = \gcd(a, b) \quad (2)$$

Extended euclidean algorithm is beneficial when a and b are coprime since x is the modular multiplicative inverse of a modulo b, and y is the modular multiplicative inverse of b modulo a. Extended Euclidean algorithm commonly appliedin cryptography [12].

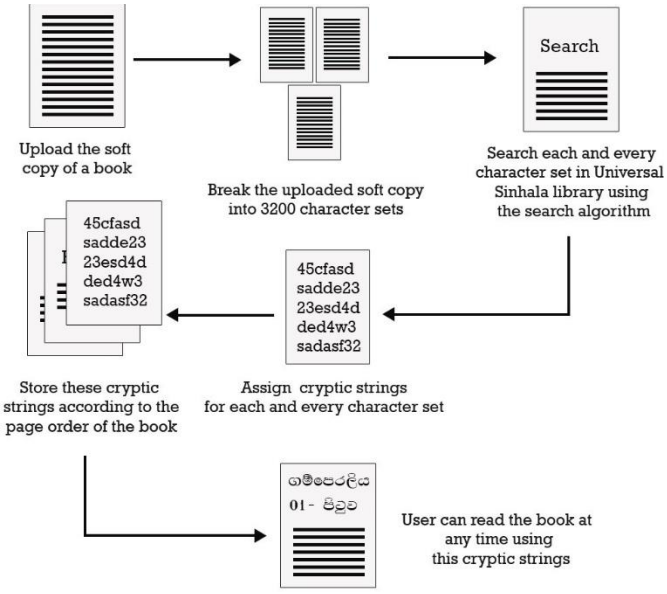

Fig. 7. Breaking down Sinhala content.

Base conversion algorithms and pseudo-random generator algorithms have been demonstrated above, leaving the third component, breaking down Sinhala content into a form that the core algorithm supports (Fig. 7). As mentioned above core algorithm takes only 3200 Sinhala characters once; therefore, Sinhala content is broken down into 3200 Sinhala character strings and merge encrypted strings at the end as the last step of the encryption algorithm. The decryption algorithm is designed to split the encrypted text into encrypted strings, convert each into Sinhala content, and merge them at the end. A user-specific character can be used when merging and splitting the strings.

Above mentioned virtual library architecture is completely customizable. The number of characters per page, the structure of the page location, the number base of the hexagon number can be changed according to the user preference. Moreover, it can be specified as the private key for the encryption method.

III. Results

As per the testing, Sinhala texts have been used to test the developed encryption algorithm. Therefore, Sinhala text should be fed to the system as a .txt file, and Sinhala context should be in Unicode characters. Furthermore, a validation process has been conducted to determine the chosen Sinhala texts necessarily represent all the characters in the Sinhala alphabet.

The above figure (Fig. 8) outlines a sample Sinhala text which has been used for encryption. After the encryption, once observed, the encrypted text is unreadable. The below figure (Fig. 9) manifests the unreadable encrypted Sinhala text file.

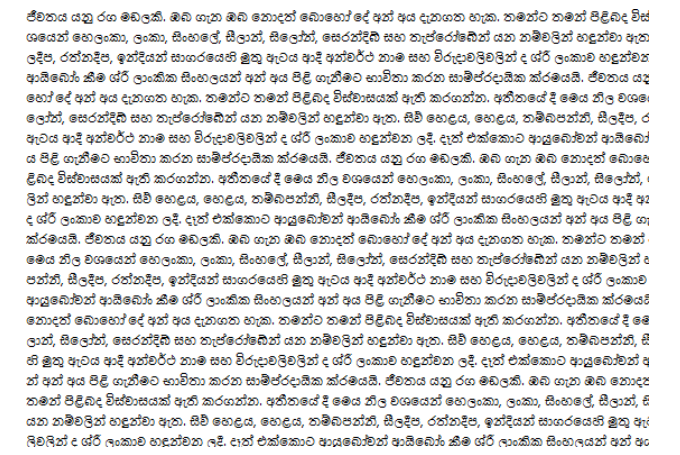

Fig. 8. Sample Sinhala text.

Fig. 9. Encrypted text.

Furthermore, the encrypted text has been fed to the decryption algorithm. After the decryption, the decrypted text was 100% similar to the original Sinhala text (Fig. 8) used for encryption. More Sinhala texts extracted from Sinhala books and documents have been used to test the Universal Sinhala Library encryption platform, and the platform successfullypassed all test cases.

IV. Conclusion

Security issues have become a significant concern for data transferring, including text sharing. This paper introduces an encryption platform designed for the Sinhala language. The architecture of the algorithm is inherited from a virtual library architecture. The core algorithm has been designed using the linear congruential generator, extended euclidean algorithm, and Hull-Dobell Therom. The core algorithm of the encryption algorithm is customizable; therefore, users can change the core algorithm preferably. As per the recommendations, this platform can be extended for any other language or any character-specific encryption. This research has the potential to be continued towards text compression and text hashing as well.

References

[1] R. L. Rivest, “The RC5 encryption algorithm”, in International Workshop on Fast Software Encryption, 1994, bll 86–96.

[2] X. Jing, Y. Hao, H. Fei, en Z. Li, “Text encryption algorithm based on natural language processing”, in 2012 Fourth International Conference on Multimedia Information Networking and Security, 2012, bll 670–672.

[3] M. A. Murillo-Escobar, F. Abundiz-Pérez, C. Cruz-Hernández, en R. M. López-Gutiérrez, “A novel symmetric text encryption algorithm based on logistic map”, in Proceedings of the international conference on communications, signal processing and computers, 2014, vol 4953.

[4] H. Kruse en A. Mukherjee, “Data compression using text encryption”, in Proceedings DCC’97. Data Compression Conference, 1997, bl 447.

[5] L. D. Singh en K. M. Singh, “Implementation of text encryption using elliptic curve cryptography”, Procedia Computer Science, vol 54, bll 73–82, 2015.

[6] U. Singh en U. Garg, “An ASCII value based text data encryption System”, International Journal of Scientific and Research Publications, vol 3, no 11, bll 2250–3153, 2013.

[7] J. L. Borges, “The Garden of Branching Paths”, Trans. Norman Thomas di Giovanni. Np: np, nd libraryofbabel. info. Web, vol 2, 2015.

[8] J. Basile, Library of Babel. 2015.

[9] D. Chandralal, Sinhala, vol 15. John Benjamins Publishing, 2010.

[10] D. De Silva et al., “Sinhala to english language translator”, in 2008 4th International Conference on Information and Automation for Sustainability, 2008, bll 419–424.

[11] Wikipedia, “Linear congruential generator — Wikipedia, The Free Encyclopedia”, 2021. [Online]. Available at: http://en.wikipedia.org/w/index.php?title=Linear

[12] Wikipedia, “Extended Euclidean algorithm — Wikipedia, The Free Encyclopedia”, 2021. [Online]. Available at: http://en.wikipedia.org/w/index.php?title=Extended